

ADVANCED MODERN ALGEBRA

JOSEPH J. ROTMAN

Advanced Modern Algebra

by Joseph J. Rotman

Hardcover: 1040 pages

Publisher: Prentice Hall; 1st edition (2002); 2nd printing (2003)

Language: English

ISBN: 0130878685

Book Description
This book's organizing principle is the interplay between groups and rings, where "rings" includes the ideas of modules. It contains basic definitions, complete and clear theorems (the first with brief sketches of proofs), and gives attention to the topics of algebraic geometry, computers, homology, and representations. More than merely a succession of definition-theorem-proofs, this text put results and ideas in context so that students can appreciate why a certain topic is being studied, and where definitions originate. Chapter topics include groups; commutative rings; modules; principal ideal domains; algebras; cohomology and representations; and homological algebra. For individuals interested in a self-study guide to learning advanced algebra and its related topics.

Book Info
Contains basic definitions, complete and clear theorems, and gives attention to the topics of algebraic geometry, computers, homology, and representations. For individuals interested in a self-study guide to learning advanced algebra and its related topics.

To my wife

Marganit

and our two wonderful kids,

Danny and Ella,

whom I love very much

Contents

Second Printing

It is my good fortune that several readers of the first printing this book apprised me of errata I had not noticed, often giving suggestions for improvement. I give special thanks to Nick Loehr, Robin Chapman, and David Leep for their generous such help.

Prentice Hall has allowed me to correct every error found; this second printing is surely better than the first one.

Joseph Rotman
May 2003

Preface

Algebra is used by virtually all mathematicians, be they analysts, combinatorists, computer scientists, geometers, logicians, number theorists, or topologists. Nowadays, everyone agrees that some knowledge of linear algebra, groups, and commutative rings is necessary, and these topics are introduced in undergraduate courses. We continue their study.

This book can be used as a text for the first year of graduate algebra, but it is much more than that. It can also serve more advanced graduate students wishing to learn topics on their own; while not reaching the frontiers, the book does provide a sense of the successes and methods arising in an area. Finally, this is a reference containing many of the standard theorems and definitions that users of algebra need to know. Thus, the book is not only an appetizer, but a hearty meal as well.

When I was a student, Birkhoff and Mac Lane's *A Survey of Modern Algebra* was the text for my first algebra course, and van der Waerden's *Modern Algebra* was the text for my second course. Both are excellent books (I have called this book *Advanced Modern Algebra* in homage to them), but times have changed since their first appearance: Birkhoff and Mac Lane's book first appeared in 1941, and van der Waerden's book first appeared in 1930. There are today major directions that either did not exist over 60 years ago, or that were not then recognized to be so important. These new directions involve algebraic geometry, computers, homology, and representations (*A Survey of Modern Algebra* has been rewritten as Mac Lane–Birkhoff, *Algebra*, Macmillan, New York, 1967, and this version introduces categorical methods; category theory emerged from algebraic topology, but was then used by Grothendieck to revolutionize algebraic geometry).

Let me now address readers and instructors who use the book as a text for a beginning graduate course. If I could assume that everyone had already read my book, *A First Course in Abstract Algebra*, then the prerequisites for this book would be plain. But this is not a realistic assumption; different undergraduate courses introducing abstract algebra abound, as do texts for these courses. For many, linear algebra concentrates on matrices and vector spaces over the real numbers, with an emphasis on computing solutions of linear systems of equations; other courses may treat vector spaces over arbitrary fields, as well as Jordan and rational canonical forms. Some courses discuss the Sylow theorems; some do not; some courses classify finite fields; some do not.

To accommodate readers having different backgrounds, the first three chapters contain

many familiar results, with many proofs merely sketched. The first chapter contains the fundamental theorem of arithmetic, congruences, De Moivre's theorem, roots of unity, cyclotomic polynomials, and some standard notions of set theory, such as equivalence relations and verification of the group axioms for symmetric groups. The next two chapters contain both familiar and unfamiliar material. "New" results, that is, results rarely taught in a first course, have complete proofs, while proofs of "old" results are usually sketched. In more detail, Chapter 2 is an introduction to group theory, reviewing permutations, Lagrange's theorem, quotient groups, the isomorphism theorems, and groups acting on sets. Chapter 3 is an introduction to commutative rings, reviewing domains, fraction fields, polynomial rings in one variable, quotient rings, isomorphism theorems, irreducible polynomials, finite fields, and some linear algebra over arbitrary fields. Readers may use "older" portions of these chapters to refresh their memory of this material (and also to see my notational choices); on the other hand, these chapters can also serve as a guide for learning what may have been omitted from an earlier course (complete proofs can be found in *A First Course in Abstract Algebra*). This format gives more freedom to an instructor, for there is a variety of choices for the starting point of a course of lectures, depending on what best fits the backgrounds of the students in a class. I expect that most instructors would begin a course somewhere in the middle of Chapter 2 and, afterwards, would continue from some point in the middle of Chapter 3. Finally, this format is convenient for the author, because it allows me to refer back to these earlier results in the midst of a discussion or a proof. Proofs in subsequent chapters are complete and are not sketched.

I have tried to write clear and complete proofs, omitting only those parts that are truly routine; thus, it is not necessary for an instructor to expound every detail in lectures, for students should be able to read the text.

Here is a more detailed account of the later chapters of this book.

Chapter 4 discusses fields, beginning with an introduction to Galois theory, the interrelationship between rings and groups. We prove the insolvability of the general polynomial of degree 5, the fundamental theorem of Galois theory, and applications, such as a proof of the fundamental theorem of algebra, and Galois's theorem that a polynomial over a field of characteristic 0 is solvable by radicals if and only if its Galois group is a solvable group.

Chapter 5 covers finite abelian groups (basis theorem and fundamental theorem), the Sylow theorems, Jordan–Hölder theorem, solvable groups, simplicity of the linear groups $\mathrm{PSL}(2, k)$, free groups, presentations, and the Nielsen–Schreier theorem (subgroups of free groups are free).

Chapter 6 introduces prime and maximal ideals in commutative rings; Gauss's theorem that $R[x]$ is a UFD when R is a UFD; Hilbert's basis theorem, applications of Zorn's lemma to commutative algebra (a proof of the equivalence of Zorn's lemma and the axiom of choice is in the appendix), inseparability, transcendence bases, Lüroth's theorem, affine varieties, including a proof of the Nullstellensatz for uncountable algebraically closed fields (the full Nullstellensatz, for varieties over arbitrary algebraically closed fields, is proved in Chapter 11); primary decomposition; Gröbner bases. Chapters 5 and 6 overlap two chapters of *A First Course in Abstract Algebra*, but these chapters are not covered in most

undergraduate courses.

Chapter 7 introduces modules over commutative rings (essentially proving that all R-modules and R-maps form an abelian category); categories and functors, including products and coproducts, pullbacks and pushouts, Grothendieck groups, inverse and direct limits, natural transformations; adjoint functors; free modules, projectives, and injectives.

Chapter 8 introduces noncommutative rings, proving Wedderburn's theorem that finite division rings are commutative, as well as the Wedderburn–Artin theorem classifying semisimple rings. Modules over noncommutative rings are discussed, along with tensor products, flat modules, and bilinear forms. We also introduce character theory, using it to prove Burnside's theorem that finite groups of order $p^m q^n$ are solvable. We then introduce multiply transitive groups and Frobenius groups, and we prove that Frobenius kernels are normal subgroups of Frobenius groups.

Chapter 9 considers finitely generated modules over PIDs (generalizing earlier theorems about finite abelian groups), and then goes on to apply these results to rational, Jordan, and Smith canonical forms for matrices over a field (the Smith normal form enables one to compute elementary divisors of a matrix). We also classify projective, injective, and flat modules over PIDs. A discussion of graded k-algebras, for k a commutative ring, leads to tensor algebras, central simple algebras and the Brauer group, exterior algebra (including Grassmann algebras and the binomial theorem), determinants, differential forms, and an introduction to Lie algebras.

Chapter 10 introduces homological methods, beginning with semidirect products and the extension problem for groups. We then present Schreier's solution of the extension problem using factor sets, culminating in the Schur–Zassenhaus lemma. This is followed by axioms characterizing Tor and Ext (existence of these functors is proved with derived functors), some cohomology of groups, a bit of crossed product algebras, and an introduction to spectral sequences.

Chapter 11 returns to commutative rings, discussing localization, integral extensions, the general Nullstellensatz (using Jacobson rings), Dedekind rings, homological dimensions, the theorem of Serre characterizing regular local rings as those noetherian local rings of finite global dimension, the theorem of Auslander and Buchsbaum that regular local rings are UFDs.

Each generation should survey algebra to make it serve the present time.

It is a pleasure to thank the following mathematicians whose suggestions have greatly improved my original manuscript: Ross Abraham, Michael Barr, Daniel Bump, Heng Huat Chan, Ulrich Daepp, Boris A. Datskovsky, Keith Dennis, Vlastimil Dlab, Sankar Dutta, David Eisenbud, E. Graham Evans, Jr., Daniel Flath, Jeremy J. Gray, Daniel Grayson, Phillip Griffith, William Haboush, Robin Hartshorne, Craig Huneke, Gerald J. Janusz, David Joyner, Carl Jockusch, David Leep, Marcin Mazur, Leon McCulloh, Emma Previato, Eric Sommers, Stephen V. Ullom, Paul Vojta, William C. Waterhouse, and Richard Weiss.

Joseph Rotman

Etymology

The heading *etymology* in the index points the reader to derivations of certain mathematical terms. For the origins of other mathematical terms, we refer the reader to my books *Journey into Mathematics* and *A First Course in Abstract Algebra*, which contain etymologies of the following terms.

Journey into Mathematics:

π, algebra, algorithm, arithmetic, completing the square, cosine, geometry, irrational number, isoperimetric, mathematics, perimeter, polar decomposition, root, scalar, secant, sine, tangent, trigonometry.

A First Course in Abstract Algebra:

affine, binomial, coefficient, coordinates, corollary, degree, factor, factorial, group, induction, Latin square, lemma, matrix, modulo, orthogonal, polynomial, quasicyclic, September, stochastic, theorem, translation.

Special Notation

$\varprojlim A_i$	inverse limit	500
$\varinjlim A_i$	direct limit	505
G'	commutator subgroup	284
G_x	stabilizer of an element x	100
$G[m]$	$\{g \in G : mg = 0\}$, where G is an additive abelian group	267
mG	$\{mg : g \in G\}$, where G is an additive abelian group	253
G_p	p-primary component of an abelian group G	256
$k[x]$	polynomials	127
$k(x)$	rational functions	129
$k[[x]]$	formal power series	130
$k\langle X\rangle$	polynomials in noncommuting variables	724
R^{op}	opposite ring	529
Ra or (a)	principal ideal generated by a	146
$R^\times$	nonzero elements in a ring R	125
$H \le G$	H is a subgroup of a group G	62
$H < G$	H is a proper subgroup of a group G	62
$H \lhd G$	H is a normal subgroup of a group G	76
$A \subseteq B$	A is a submodule (subring) of a module (ring)B	119
$A \subsetneq B$	A is a proper submodule (subring) of a module (ring)B	119
1_X	identity function on a set X	
1_X	identity morphism on an object X	443
$f : a \mapsto b$	$f(a) = b$	28
$\lvert X\rvert$	number of elements in a set X	
${}_Y[T]_X$	matrix of a linear transformation T relative to bases X and Y	173
χ_σ	character afforded by a representation σ	610
$\phi(n)$	Euler ϕ-function	21
$\binom{n}{r}$	binomial coefficient $n!/r!(n-r)!$	5
δ_{ij}	Kronecker delta $\delta_{ij} = \begin{cases} 1 & \text{if } i = j; \\ 0 & \text{if } i \ne j. \end{cases}$	
$a_1, \ldots, \widehat{a_i}, \ldots, a_n$	list $a_1, \ldots, a_n$ with a_i omitted	

1

Things Past

This chapter reviews some familiar material of number theory, complex roots of unity, and basic set theory, and so most proofs are merely sketched.

1.1 Some Number Theory

Let us begin by discussing mathematical induction. Recall that the set of ***natural numbers*** $\mathbb{N}$ is defined by

$$\mathbb{N} = \{\text{integers } n : n \geq 0\};$$

that is, $\mathbb{N}$ is the set of all nonnegative integers. Mathematical induction is a technique of proof based on the following property of $\mathbb{N}$:

Least Integer Axiom.[1] There is a smallest integer in every nonempty subset C of $\mathbb{N}$.

Assuming the axiom, let us see that if m is any fixed integer, possibly negative, then there is a smallest integer in every nonempty collection C of integers greater than or equal to m. If $m \geq 0$, this is the least integer axiom. If $m < 0$, then $C \subseteq \{m, m+1, \ldots, -1\} \cup \mathbb{N}$ and

$$C = \big(C \cap \{m, m+1, \ldots, -1\}\big) \cup \big(C \cap \mathbb{N}\big).$$

If the finite set $C \cap \{m, m+1, \ldots, -1\} \neq \varnothing$, then it contains a smallest integer that is, obviously, the smallest integer in C; if $C \cap \{m, m+1, \ldots, -1\} = \varnothing$, then C is contained in $\mathbb{N}$, and the least integer axiom provides a smallest integer in C.

Definition. A natural number p is ***prime*** if $p \geq 2$ and there is no factorization $p = ab$, where $a < p$ and $b < p$ are natural numbers.

[1] This property is usually called the *well-ordering principle*.

Proposition 1.1. *Every integer $n \geq 2$ is either a prime or a product of primes.*

Proof. Let C be the subset of $\mathbb{N}$ consisting of all those $n \geq 2$ for which the proposition is false; we must prove that $C = \varnothing$. If, on the contrary, C is nonempty, then it contains a smallest integer, say, m. Since $m \in C$, it is not a prime, and so there are natural numbers a and b with $m = ab$, $a < m$, and $b < m$. Neither a nor b lies in C, for each of them is smaller than m, which is the smallest integer in C, and so each of them is either prime or a product of primes. Therefore, $m = ab$ is a product of (at least two) primes, contradicting the proposition being false for m. •

There are two versions of induction.

Theorem 1.2 (Mathematical Induction). *Let $S(n)$ be a family of statements, one for each integer $n \geq m$, where m is some fixed integer. If*

(i) *$S(m)$ is true, and*

(ii) *$S(n)$ is true implies $S(n+1)$ is true,*

then $S(n)$ is true for all integers $n \geq m$.

Proof. Let C be the set of all integers $n \geq m$ for which $S(n)$ is false. If C is empty, we are done. Otherwise, there is a smallest integer k in C. By (i), we have $k > m$, and so there is a statement $S(k-1)$. But $k - 1 < k$ implies $k - 1 \notin C$, for k is the smallest integer in C. Thus, $S(k-1)$ is true. But now (ii) says that $S(k) = S([k-1]+1)$ is true, and this contradicts $k \in C$ [which says that $S(k)$ is false]. •

Theorem 1.3 (Second Form of Induction). *Let $S(n)$ be a family of statements, one for each integer $n \geq m$, where m is some fixed integer. If*

(i) *$S(m)$ is true, and*

(ii) *if $S(k)$ is true for all k with $m \leq k < n$, then $S(n)$ is itself true,*

then $S(n)$ is true for all integers $n \geq m$.

Sketch of Proof. The proof is similar to the proof of the first form. •

We now recall some elementary number theory.

Theorem 1.4 (Division Algorithm). *Given integers a and b with $a \neq 0$, there exist unique integers q and r with*

$$b = qa + r \quad \text{and} \quad 0 \leq r < |a|.$$

Sketch of Proof. Consider all nonnegative integers of the form $b - na$, where $n \in \mathbb{Z}$. Define r to be the smallest nonnegative integer of the form $b - na$, and define q to be the integer n occurring in the expression $r = b - na$.

If $qa + r = q'a + r'$, where $0 \leq r' < |a|$, then $|(q - q')a| = |r' - r|$. Now $0 \leq |r' - r| < |a|$ and, if $|q - q'| \neq 0$, then $|(q - q')a| \geq |a|$. We conclude that both sides are 0; that is, $q = q'$ and $r = r'$. •

Definition. If a and b are integers with $a \neq 0$, then the integers q and r occurring in the division algorithm are called the ***quotient*** and the ***remainder*** after dividing b by a.

Warning! The division algorithm makes sense, in particular, when b is negative. A careless person may assume that b and $-b$ leave the same remainder after dividing by a, and this is usually false. For example, let us divide 60 and -60 by 7.

$$60 = 7 \cdot 8 + 4 \text{ and } -60 = 7 \cdot (-9) + 3$$

Thus, the remainders after dividing 60 and -60 by 7 are different.

Corollary 1.5. *There are infinitely many primes.*

*Proof. (**Euclid**)* Suppose, on the contrary, that there are only finitely many primes. If $p_1, p_2, \ldots, p_k$ is the complete list of all the primes, define $M = (p_1 \cdots p_k) + 1$. By Proposition 1.1, M is either a prime or a product of primes. But M is neither a prime ($M > p_i$ for every i) nor does it have any prime divisor p_i, for dividing M by p_i gives remainder 1 and not 0. For example, dividing M by p_1 gives $M = p_1(p_2 \cdots p_k) + 1$, so that the quotient and remainder are $q = p_2 \cdots p_k$ and $r = 1$; dividing M by p_2 gives $M = p_2(p_1 p_3 \cdots p_k) + 1$, so that $q = p_1 p_3 \cdots p_k$ and $r = 1$; and so forth. This contradiction proves that there cannot be only finitely many primes, and so there must be an infinite number of them. •

Definition. If a and b are integers, then a is a ***divisor*** of b if there is an integer d with $b = ad$. We also say that a ***divides*** b or that b is a ***multiple*** of a, and we denote this by

$$a \mid b.$$

There is going to be a shift in viewpoint. When we first learned long division, we emphasized the quotient q; the remainder r was merely the fragment left over. Here, we are interested in whether or not a given number b is a multiple of a number a, but we are less interested in which multiple it may be. Hence, from now on, we will emphasize the remainder. Thus, $a \mid b$ if and only if b has remainder $r = 0$ after dividing by a.

Definition. A ***common divisor*** of integers a and b is an integer c with $c \mid a$ and $c \mid b$. The ***greatest common divisor***or **gcd** of a and b, denoted by (a, b), is defined by

$$(a, b) = \begin{cases} 0 \text{ if } a = 0 = b \\ \text{the largest common divisor of } a \text{ and } b \text{ otherwise.} \end{cases}$$

Proposition 1.6. *If p is a prime and b is any integer, then*

$$(p, b) = \begin{cases} p & \text{if } p \mid b \\ 1 & \text{otherwise.} \end{cases}$$

Sketch of Proof. A positive common divisor is, in particular, a divisor of the prime p, and hence it is p or 1. •

Theorem 1.7. *If a and b are integers, then $(a, b) = d$ is a linear combination of a and b; that is, there are integers s and t with $d = sa + tb$.*

Sketch of Proof. Let

$$I = \{sa + tb : s, t \in \mathbb{Z}\}$$

(the set of all integers, positive and negative, is denoted by $\mathbb{Z}$). If $I \neq \{0\}$, let d be the smallest positive integer in I; as any element of I, we have $d = sa + tb$ for some integers s and t. We claim that $I = (d)$, the set of all multiples of d. Clearly, $(d) \subseteq I$. For the reverse inclusion, take $c \in I$. By the division algorithm, $c = qd + r$, where $0 \le r < d$. Now $r = c - qd \in I$, so that the minimality of d is contradicted if $r \neq 0$. Hence, $d \mid c$, $c \in (d)$, and $I = (d)$. It follows that d is a common divisor of a and b, and it is the largest such. •

Proposition 1.8. *Let a and b be integers. A nonnegative common divisor d is their* gcd *if and only if $c \mid d$ for every common divisor c.*

Sketch of Proof. If d is the gcd, then $d = sa + tb$. Hence, if $c \mid a$ and $c \mid b$, then c divides $sa + tb = d$. Conversely, if d is a common divisor with $c \mid d$ for every common divisor c, then $c \le d$ for all c, and so d is the largest. •

Corollary 1.9. *Let I be a subset of $\mathbb{Z}$ such that*

(i) $0 \in I$;

(ii) *if $a, b \in I$, then $a - b \in I$;*

(iii) *if $a \in I$ and $q \in \mathbb{Z}$, then $qa \in I$.*

Then there is a natural number $d \in I$ with I consisting precisely of all the multiples of d.

Sketch of Proof. These are the only properties of the subset I in Theorem 1.7 that were used in the proof. •

Theorem 1.10 (Euclid's Lemma). *If p is a prime and $p \mid ab$, then $p \mid a$ or $p \mid b$. More generally, if a prime p divides a product $a_1 a_2 \cdots a_n$, then it must divide at least one of the factors a_i.*

Sketch of Proof. If $p \nmid a$, then $(p, a) = 1$ and $1 = sp + ta$. Hence, $b = spb + tab$ is a multiple of p. The second statement is proved by induction on $n \ge 2$. •

Definition. Call integers a and b ***relatively prime*** if their gcd $(a, b) = 1$.

Corollary 1.11. *Let a, b, and c be integers. If c and a are relatively prime and if $c \mid ab$, then $c \mid b$.*

Sketch of Proof. Since $1 = sc + ta$, we have $b = scb + tab$. •

Proposition 1.12. *If p is a prime, then $p \mid \binom{p}{j}$ for $0 < j < p$.*

Sketch of Proof. By definition, the binomial coefficient $\binom{p}{j} = p!/j!(p-j)!$, so that

$$p! = j!(p-j)!\binom{p}{j}.$$

By Euclid's lemma, $p \nmid j!(p-j)!$ implies $p \mid \binom{p}{j}$. •

If integers a and b are not both 0, Theorem 1.7 identifies (a, b) as the smallest positive linear combination of a and b. Usually, this is not helpful in actually finding the gcd, but the next elementary result is an exception.

Proposition 1.13.

(i) *If a and b are integers, then a and b are relatively prime if and only if there are integers s and t with $1 = sa + tb$.*

(ii) *If $d = (a, b)$, where a and b are not both 0, then $(a/d, b/d) = 1$.*

Proof. (i) Necessity is Theorem 1.7. For sufficiency, note that 1 being the smallest positive integer gives, in this case, 1 being the smallest positive linear combination of a and b, and hence $(a, b) = 1$. Alternatively, if c is a common divisor of a and b, then $c \mid sa + tb$; hence, $c \mid 1$, and so $c = \pm 1$.

(ii) Note that $d \neq 0$ and a/d and b/d are integers, for d is a common divisor. The equation $d = sa + tb$ now gives $1 = s(a/d) + t(b/d)$. By part (i), $(a/d, b/d) = 1$. •

The next result offers a practical method for finding the gcd of two integers as well as for expressing it as a linear combination.

Theorem 1.14 (Euclidean Algorithm). *Let a and b be positive integers. There is an algorithm that finds the* gcd, *$d = (a, b)$, and there is an algorithm that finds a pair of integers s and t with $d = sa + tb$.*

Remark. More details can be found in Theorem 3.40, where this result is proved for polynomials.

To see how the Greeks discovered this result, see the discussion of *antanairesis* in Rotman, *A First Course in Abstract Algebra*, page 49. ◂

Sketch of Proof. This algorithm iterates the division algorithm, as follows. Begin with $b = qa + r$, where $0 \leq r < a$. The second step is $a = q'r + r'$, where $0 \leq r' < r$; the next step is $r = q''r' + r''$, where $0 \leq r'' < r'$, and so forth. This iteration stops eventually, and the last remainder is the gcd. Working upward from the last equation, we can write the gcd as a linear combination of a and b. •

Proposition 1.15. *If $b \geq 2$ is an integer, then every positive integer m has an expression in **base b**: There are integers d_i with $0 \leq d_i < b$ such that*

$$m = d_k b^k + d_{k-1} b^{k-1} + \cdots + d_0;$$

moreover, this expression is unique if $d_k \neq 0$.

Sketch of Proof. By the least integer axiom, there is an integer $k \geq 0$ with $b^k \leq m < b^{k+1}$, and the division algorithm gives $m = d_k b^k + r$, where $0 \leq r < b^k$. The existence of b-adic digits follows by induction on $m \geq 1$. Uniqueness can also be proved by induction on m, but one must take care to treat all possible cases that may arise. •

The numbers $d_k, d_{k-1}, \ldots, d_0$ are called the ***b-adic digits*** of m.

Theorem 1.16 (Fundamental Theorem of Arithmetic). *Assume that an integer $a \geq 2$ has factorizations*

$$a = p_1 \cdots p_m \ \text{ and } \ a = q_1 \cdots q_n,$$

where the p's and q's are primes. Then $n = m$ and the q's may be reindexed so that $q_i = p_i$ for all i. Hence, there are unique distinct primes p_i and unique integers $e_i > 0$ with

$$a = p_1^{e_1} \cdots p_n^{e_n}.$$

Proof. We prove the theorem by induction on ℓ, the larger of m and n.

If $\ell = 1$, then the given equation is $a = p_1 = q_1$, and the result is obvious. For the inductive step, note that the equation gives $p_m \mid q_1 \cdots q_n$. By Euclid's lemma, there is some i with $p_m \mid q_i$. But q_i, being a prime, has no positive divisors other than 1 and itself, so that $q_i = p_m$. Reindexing, we may assume that $q_n = p_m$. Canceling, we have $p_1 \cdots p_{m-1} = q_1 \cdots q_{n-1}$. By the inductive hypothesis, $n - 1 = m - 1$ and the q's may be reindexed so that $q_i = p_i$ for all i. •

Definition. A ***common multiple*** of integers a and b is an integer c with $a \mid c$ and $b \mid c$. The ***least common multiple*** or **lcm** of a and b, denoted by $[a, b]$, is defined by

$$[a, b] = \begin{cases} 0 \text{ if } a = 0 = b \\ \text{the smallest positive common multiple of } a \text{ and } b \text{ otherwise.} \end{cases}$$

Proposition 1.17. *Let $a = p_1^{e_1} \cdots p_n^{e_n}$ and let $b = p_1^{f_1} \cdots p_n^{f_n}$, where $e_i \geq 0$ and $f_i \geq 0$ for all i; define*

$$m_i = \min\{e_i, f_i\} \quad \text{and} \quad M_i = \max\{e_i, f_i\}.$$

Then the gcd *and the* lcm *of a and b are given by*

$$(a, b) = p_1^{m_1} \cdots p_n^{m_n} \quad \text{and} \quad [a, b] = p_1^{M_1} \cdots p_n^{M_n}.$$

Sketch of Proof. Use the fact that $p_1^{e_1} \cdots p_n^{e_n} \mid p_1^{f_1} \cdots p_n^{f_n}$ if and only if $e_i \leq f_i$ for all i. •

Definition. Let $m \geq 0$ be fixed. Then integers a and b are ***congruent modulo*** m, denoted by

$$a \equiv b \bmod m,$$

if $m \mid (a - b)$.

Proposition 1.18. *If $m \geq 0$ is a fixed integer, then for all integers a, b, c,*

(i) $a \equiv a \bmod m$;

(ii) *if $a \equiv b \bmod m$, then $b \equiv a \bmod m$*;

(iii) *if $a \equiv b \bmod m$ and $b \equiv c \bmod m$, then $a \equiv c \bmod m$.*

Remark. (i) says that congruence is ***reflexive,*** (ii) says it is ***symmetric,*** and (iii) says it is ***transitive.*** ◄

Sketch of Proof. All the items follow easily from the definition of congruence. •

Proposition 1.19. *Let $m \geq 0$ be a fixed integer.*

(i) *If $a = qm + r$, then $a \equiv r \bmod m$.*

(ii) *If $0 \leq r' < r < m$, then $r \not\equiv r' \bmod m$; that is, r and r' are not congruent* mod *m.*

(iii) *$a \equiv b \bmod m$ if and only if a and b leave the same remainder after dividing by m.*

(iv) *If $m \geq 2$, each integer a is congruent* mod *m to exactly one of $0, 1, \ldots, m - 1$.*

Sketch of Proof. Items (i) and (iii) are routine; item (ii) follows after noting that $0 < r - r' < m$, and item (iv) follows from (i) and (ii). •

The next result shows that congruence is compatible with addition and multiplication.

Proposition 1.20. *Let $m \geq 0$ be a fixed integer.*

(i) *If $a \equiv a' \bmod m$ and $b \equiv b' \bmod m$, then*

$$a + b \equiv a' + b' \bmod m.$$

(ii) *If $a \equiv a' \bmod m$ and $b \equiv b' \bmod m$, then*

$$ab \equiv a'b' \bmod m.$$

(iii) *If $a \equiv b \bmod m$, then $a^n \equiv b^n \bmod m$ for all $n \geq 1$.*

Sketch of Proof. All the items are routine. •

Earlier we divided 60 and -60 by 7, getting remainders 4 in the first case and 3 in the second. It is no accident that $4+3=7$. If a is an integer and $m \geq 0$, let $a \equiv r \bmod m$ and $-a \equiv r' \bmod m$. It follows from the proposition that

$$0 = -a + a \equiv r' + r \bmod m.$$

The next example shows how one can use congruences. In each case, the key idea is to solve a problem by replacing numbers by their remainders.

Example 1.21.
(i) Prove that if a is in $\mathbb{Z}$, then $a^2 \equiv 0$, 1, or 4 mod 8.

If a is an integer, then $a \equiv r \bmod 8$, where $0 \leq r \leq 7$; moreover, by Proposition 1.20(iii), $a^2 \equiv r^2 \bmod 8$, and so it suffices to look at the squares of the remainders.

r	0	1	2	3	4	5	6	7
r^2	0	1	4	9	16	25	36	49
r^2 mod 8	0	1	4	1	0	1	4	1

Table 1.1. Squares mod 8

We see in Table 1.1 that only 0, 1, or 4 can be a remainder after dividing a perfect square by 8.

(ii) Prove that $n = 1003456789$ is not a perfect square.

Since $1000 = 8 \cdot 125$, we have $1000 \equiv 0 \bmod 8$, and so

$$n = 1003456789 = 1003456 \cdot 1000 + 789 \equiv 789 \bmod 8.$$

Dividing 789 by 8 leaves remainder 5; that is, $n \equiv 5 \bmod 8$. Were n a perfect square, then $n \equiv 0$, 1, or 4 mod 8.

(iii) If m and n are positive integers, are there any perfect squares of the form $3^m + 3^n + 1$?

Again, let us look at remainders mod 8. Now $3^2 = 9 \equiv 1 \bmod 8$, and so we can evaluate $3^m \bmod 8$ as follows: If $m = 2k$, then $3^m = 3^{2k} = 9^k \equiv 1 \bmod 8$; if $m = 2k+1$, then $3^m = 3^{2k+1} = 9^k \cdot 3 \equiv 3 \bmod 8$. Thus,

$$3^m \equiv \begin{cases} 1 \bmod 8 & \text{if } m \text{ is even;} \\ 3 \bmod 8 & \text{if } m \text{ is odd.} \end{cases}$$

Replacing numbers by their remainders after dividing by 8, we have the following possibilities for the remainder of $3^m + 3^n + 1$, depending on the parities of m and n:

$$3 + 1 + 1 \equiv 5 \bmod 8$$
$$3 + 3 + 1 \equiv 7 \bmod 8$$
$$1 + 1 + 1 \equiv 3 \bmod 8$$
$$1 + 3 + 1 \equiv 5 \bmod 8.$$

In no case is the remainder 0, 1, or 4, and so no number of the form $3^m + 3^n + 1$ can be a perfect square, by part (i). ◂

Proposition 1.22. *A positive integer a is divisible by* 3 (*or by* 9) *if and only if the sum of its* (*decimal*) *digits is divisible by* 3 (*or by* 9).

Sketch of Proof. Observe that $10^n \equiv 1 \bmod 3$ (and also that $10^n \equiv 1 \bmod 9$). •

Proposition 1.23. *If p is a prime and a and b are integers, then*

$$(a+b)^p \equiv a^p + b^p \bmod p.$$

Sketch of Proof. Use the binomial theorem and Proposition 1.12. •

Theorem 1.24 (*Fermat*). *If p is a prime, then*

$$a^p \equiv a \bmod p$$

for every a in $\mathbb{Z}$. *More generally, for every integer* $k \geq 1$,

$$a^{p^k} \equiv a \bmod p.$$

Sketch of Proof. If $a \geq 0$, use induction on a; the inductive step uses Proposition 1.23. The second statement follows by induction on $k \geq 1$. •

Corollary 1.25. *Let p be a prime and let n be a positive integer. If* $m \geq 0$ *and if* Σ *is the sum of the p-adic digits of m, then*

$$n^m \equiv n^\Sigma \bmod p.$$

Sketch of Proof. Write m in base p, and use Fermat's theorem. •

We compute the remainder after dividing 10^{100} by 7. First, $10^{100} \equiv 3^{100} \bmod 7$. Second, since $100 = 2 \cdot 7^2 + 2$, the corollary gives $3^{100} \equiv 3^4 = 81 \bmod 7$. Since $81 = 11 \times 7 + 4$, we conclude that the remainder is 4.

Theorem 1.26. *If* $(a, m) = 1$, *then, for every integer b, the congruence*

$$ax \equiv b \bmod m$$

can be solved for x; *in fact,* $x = sb$, *where* $sa \equiv 1 \bmod m$ *is one solution. Moreover, any two solutions are congruent* mod m.

Sketch of Proof. If $1 = sa + tm$, then $b = sab + tmb$. Hence, $b \equiv a(sb) \bmod m$. If, also, $b \equiv ax \bmod m$, then $0 \equiv a(x - sb) \bmod m$, so that $m \mid a(x - sb)$. Since $(m, a) = 1$, we have $m \mid (x - sb)$; hence, $x \equiv sb \bmod m$, by Corollary 1.11. •

Corollary 1.27. *If p is a prime and a is not divisible by p, then the congruence*

$$ax \equiv b \bmod p$$

is always solvable.

Sketch of Proof. If a is not divisible by p, then $(a, p) = 1$. •

Theorem 1.28 (Chinese Remainder Theorem). *If m and m' are relatively prime, then the two congruences*

$$x \equiv b \bmod m$$
$$x \equiv b' \bmod m'$$

have a common solution, and any two solutions are congruent mod mm'.

Sketch of Proof. By Theorem 1.26, any solution x to the first congruence has the form $x = sb + km$ for some $k \in \mathbb{Z}$ (where $1 = sa + tm$). Substitute this into the second congruence and solve for k. Alternatively, there are integers s and s' with $1 = sm + s'm'$, and a common solution is

$$x = b'ms + bm's'.$$

To prove uniqueness, assume that $y \equiv b \bmod m$ and $y \equiv b' \bmod m'$. Then $x - y \equiv 0 \bmod m$ and $x - y \equiv 0 \bmod m'$; that is, both m and m' divide $x - y$. The result now follows from Exercise 1.19 on page 13. •

EXERCISES

1.1 Prove that $1^2 + 2^2 + \cdots + n^2 = \frac{1}{6}n(n+1)(2n+1) = \frac{1}{3}n^3 + \frac{1}{2}n^2 + \frac{1}{6}n$.

1.2 Prove that $1^3 + 2^3 + \cdots + n^3 = \frac{1}{4}n^4 + \frac{1}{2}n^3 + \frac{1}{4}n^2$.

1.3 Prove that $1^4 + 2^4 + \cdots + n^4 = \frac{1}{5}n^5 + \frac{1}{2}n^4 + \frac{1}{3}n^3 - \frac{1}{30}n$.

Remark. There is a general formula that expresses $\sum_{i=1}^{n-1} i^k$, for $k \geq 1$, as a polynomial in n:

$$(k+1)\sum_{i=1}^{n-1} i^k = n^{k+1} + \sum_{j=1}^{k} \binom{k+1}{j} B_j n^{k+1-j};$$

the coefficients involve rational numbers B_j, for $j \geq 1$, called ***Bernoulli numbers***, defined by

$$\frac{x}{e^x - 1} = 1 + \sum_{j \geq 1} \frac{B_j}{j!} x^j;$$

see Borevich–Shafarevich, *Number Theory*, page 382. ◄

1.4 Derive the formula for $\sum_{i=1}^{n} i$ by computing the area $(n+1)^2$ of a square with sides of length $n+1$ using Figure 1.1.
Hint. The triangular areas on either side of the diagonal have equal area.

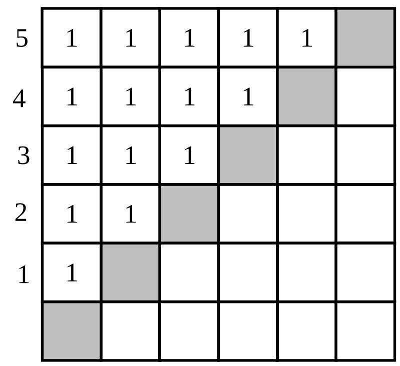


Figure 1.1

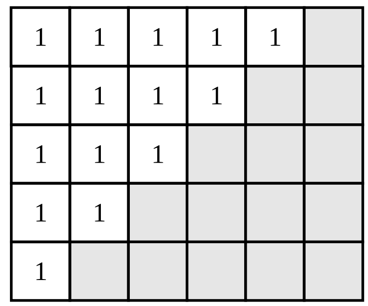


Figure 1.2

1.5 (i) Derive the formula for $\sum_{i=1}^{n} i$ by computing the area $n(n+1)$ of a rectangle with base $n+1$ and height n, as pictured in Figure 1.2.

(ii) (***Alhazen***, ca. 965-1039) For fixed $k \geq 1$, use Figure 1.3 to prove

$$(n+1)\sum_{i=1}^{n} i^k = \sum_{i=1}^{n} i^{k+1} + \sum_{i=1}^{n}\Big(\sum_{\ell=1}^{i} \ell^k\Big).$$

Hint. As indicated in Figure 1.3, a rectangle with height $n+1$ and base $\sum_{i=1}^{n} i^k$ can be subdivided so that the shaded staircase has area $\sum_{i=1}^{n} i^{k+1}$, whereas the area above it is

$$1^k + (1^k + 2^k) + (1^k + 2^k + 3^k) + \cdots + (1^k + 2^k + \cdots + n^k).$$

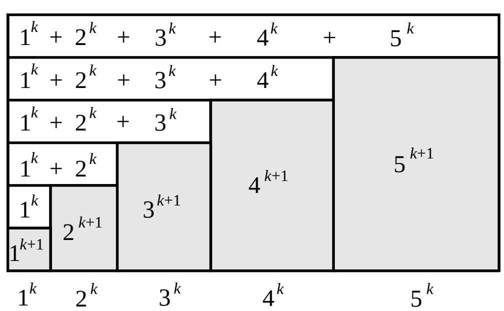


Figure 1.3

(iii) Given the formula $\sum_{i=1}^{n} = \frac{1}{2}n(n+1)$, use part (ii) to derive the formula for $\sum_{i=1}^{n} i^2$.

Hint. In Alhazen's formula, write $\sum_{i=1}^{n}\left(\sum_{\ell=1}^{i} \ell\right) = \frac{1}{2}\sum_{i=1}^{n} i^2 + \frac{1}{2}\sum_{i=1}^{n} i$, and then solve for $\sum_{i=1}^{n} i^2$ in terms of the rest.

1.6 (***Leibniz***) A function $f : \mathbb{R} \to \mathbb{R}$ is called a C^∞***-function*** if it has an nth derivative $f^{(n)}$ for every natural number n ($f^{(0)}$ is defined to be f). If f and g are C^∞-functions, prove that

$$(fg)^{(n)} = \sum_{r=0}^{n} \binom{n}{r} f^{(r)} \cdot g^{(n-r)}.$$

1.7 (**Double Induction**) Let $S(m, n)$ be a doubly indexed family of statements, one for each $m \geq 1$ and $n \geq 1$. Suppose that

(i) $S(1, 1)$ is true;

(ii) if $S(m, 1)$ is true, then $S(m + 1, 1)$ is true;

(iii) if $S(m, n)$ is true for all m, then $S(m, n + 1)$ is true for all m.

Prove that $S(m, n)$ is true for all $m \geq 1$ and $n \geq 1$.

1.8 Use double induction to prove that

$$(m+1)^n > mn$$

for all $m, n \geq 1$.

1.9 Prove that $\sqrt{2}$ is irrational.
Hint. If $\sqrt{2}$ is rational, then $\sqrt{2} = a/b$, and we can assume that $(a, b) = 1$ (actually, it is enough to assume that at least one of a and b is odd). Squaring this equation leads to a contradiction.

1.10 Prove the converse of Euclid's lemma: An integer $p \geq 2$, which, whenever it divides a product necessarily divides one of the factors, must be a prime.

1.11 Let $p_1, p_2, p_3, \ldots$ be the list of the primes in ascending order: $p_1 = 2$, $p_2 = 3$, $p_3 = 5, \ldots$ Define $f_k = p_1 p_2 \cdots p_k + 1$ for $k \geq 1$. Find the smallest k for which f_k is not a prime.
Hint. $19 \mid f_7$, but 7 is not the smallest k.

1.12 If d and d' are nonzero integers, each of which divides the other, prove that $d' = \pm d$.

1.13 Show that every positive integer m can be written as a sum of distinct powers of 2; show, moreover, that there is only one way in which m can so be written.
Hint. Write m in base 2.

1.14 If $(r, a) = 1 = (r', a)$, prove that $(rr', a) = 1$.

1.15 (i) Prove that if a positive integer n is ***squarefree*** (i.e., n is not divisible by the square of any prime), then $\sqrt{n}$ is irrational.

(ii) Prove that an integer $m \geq 2$ is a perfect square if and only if each of its prime factors occurs an even number of times.

1.16 Prove that $\sqrt[3]{2}$ is irrational.
Hint. Assume that $\sqrt[3]{2}$ can be written as a fraction in lowest terms.

1.17 Find the gcd $d = (12327, 2409)$, find integers s and t with $d = 12327s + 2409t$, and put the fraction 2409/12327 in lowest terms.

1.18 Assume that $d = sa + tb$ is a linear combination of integers a and b. Find infinitely many pairs of integers (s_k, t_k) with
$$d = s_k a + t_k b.$$
Hint. If $2s + 3t = 1$, then $2(s + 3) + 3(t - 2) = 1$.

1.19 If a and b are relatively prime and if each divides an integer n, then their product ab also divides n.

1.20 If $a > 0$, prove that $a(b, c) = (ab, ac)$. [We must assume that $a > 0$ lest $a(b, c)$ be negative.]
Hint. Show that if k is a common divisor of ab and ac, then $k \mid a(b, c)$.

Definition. A ***common divisor*** of integers $a_1, a_2, \ldots, a_n$ is an integer c with $c \mid a_i$ for all i; the largest of the common divisors, denoted by $(a_1, a_2, \ldots, a_n)$, is called the ***greatest common divisor***.

1.21 (i) Show that if d is the greatest common divisor of $a_1, a_2, \ldots, a_n$, then $d = \sum t_i a_i$, where t_i is in $\mathbb{Z}$ for $1 \le i \le n$.
(ii) Prove that if c is a common divisor of $a_1, a_2, \ldots, a_n$, then $c \mid d$.

1.22 (i) Show that (a, b, c), the gcd of a, b, c, is equal to $(a, (b, c))$.
(ii) Compute (120, 168, 328).

1.23 A ***Pythagorean triple*** is an ordered triple (a, b, c) of positive integers for which
$$a^2 + b^2 = c^2;$$
it is called ***primitive*** if gcd $(a, b, c) = 1$.
(i) If $q > p$ are positive integers, prove that
$$(q^2 - p^2, 2qp, q^2 + p^2)$$
is a Pythagorean triple. [One can prove that every *primitive* Pythagorean triple (a, b, c) is of this type.]
(ii) Show that the Pythagorean triple (9, 12, 15) (which is not primitive) is not of the type given in part (i).
(iii) Using a calculator that can find square roots but that can display only 8 digits, prove that
$$(19597501, 28397460, 34503301)$$
is a Pythagorean triple by finding q and p.

Definition. A ***common multiple*** of $a_1, a_2, \ldots, a_n$ is an integer m with $a_i \mid m$ for all i. The ***least common multiple***, written lcm and denoted by $[a_1, a_2, \ldots, a_n]$, is the smallest positive common multiple if all $a_i \ne 0$, and it is 0 otherwise.

1.24 Prove that an integer $M \ge 0$ is the lcm of $a_1, a_2, \ldots, a_n$ if and only if it is a common multiple of $a_1, a_2, \ldots, a_n$ that divides every other common multiple.

1.25 Let $a_1/b_1, \ldots, a_n/b_n \in \mathbb{Q}$, where $(a_i, b_i) = 1$ for all i. If $M = \text{lcm}\{b_1, \ldots, b_n\}$, prove that the gcd of $Ma_1/b_1, \ldots, Ma_n/b_n$ is 1.

1.26 (i) Prove that $[a, b](a, b) = ab$, where $[a, b]$ is the least common multiple of a and b.
Hint. If neither a nor b is 0, show that $ab/(a, b)$ is a common multiple of a and b that divides every common multiple c of a and b. Alternatively, use Proposition 1.17.

1.27 (i) Find the gcd (210, 48) using factorizations into primes.

(ii) Find $(1234, 5678)$.

1.28 If a and b are positive integers with $(a, b) = 1$, and if ab is a square, prove that both a and b are squares.

Hint. The sets of prime divisors of a and b are disjoint.

1.29 Let $n = p^r m$, where p is a prime not dividing an integer $m \geq 1$. Prove that

$$p \nmid \binom{n}{p^r}.$$

Hint. Assume otherwise, cross multiply, and use Euclid's lemma.

1.30 Let m be a positive integer, and let m' be an integer obtained from m by rearranging its (decimal) digits (e.g., take $m = 314159$ and $m' = 539114$). Prove that $m - m'$ is a multiple of 9.

1.31 Prove that a positive integer n is divisible by 11 if and only if the alternating sum of its digits is divisible by 11 (if the digits of a are $d_k \ldots d_2 d_1 d_0$, then their ***alternating sum*** is $d_0 - d_1 + d_2 - \cdots$).

Hint. $10 \equiv -1 \mod 11$.

1.32 (i) Prove that $10q + r$ is divisible by 7 if and only if $q - 2r$ is divisible by 7.

(ii) Given an integer a with decimal digits $d_k d_{k-1} \ldots d_0$, define

$$a' = d_k d_{k-1} \cdots d_1 - 2d_0.$$

Show that a is divisible by 7 if and only if some one of a', a'', a''',... is divisible by 7. (For example, if $a = 65464$, then $a' = 6546 - 8 = 6538$, $a'' = 653 - 16 = 637$, and $a''' = 63 - 14 = 49$; we conclude that 65464 is divisible by 7.)

1.33 (i) Show that $1000 \equiv -1 \mod 7$.

(ii) Show that if $a = r_0 + 1000r_1 + 1000^2 r_2 + \cdots$, then a is divisible by 7 if and only if $r_0 - r_1 + r_2 - \cdots$ is divisible by 7.

Remark. Exercises 1.32 and 1.33 combine to give an efficient way to determine whether large numbers are divisible by 7. If $a = 33456789123987$, for example, then $a \equiv 0 \mod 7$ if and only if $987 - 123 + 789 - 456 + 33 = 1230 \equiv 0 \mod 7$. By Exercise 1.32, $1230 \equiv 123 \equiv 6 \mod 7$, and so a is not divisible by 7. ◄

1.34 Prove that there are no integers x, y, and z such that

$$x^2 + y^2 + z^2 = 999.$$

Hint. Use Example 1.21(i).

1.35 Prove that there is no perfect square a^2 whose last two digits are 35.

Hint. If the last digit of a^2 is 5, then $a^2 \equiv 5 \mod 10$; if the last two digits of a^2 are 35, then $a^2 \equiv 35 \mod 100$.

1.36 If x is an odd number not divisible by 3, prove that $x^2 \equiv 1 \mod 4$.

1.37 Prove that if p is a prime and if $a^2 \equiv 1 \mod p$, then $a \equiv \pm 1 \mod p$.

Hint. Use Euclid's lemma.

1.38 If $(a, m) = d$, prove that $ax \equiv b \bmod m$ has a solution if and only if $d \mid b$.

1.39 Solve the congruence $x^2 \equiv 1 \bmod 21$.
Hint. Use Euclid's lemma with $21 \mid (a+1)(a-1)$.

1.40 Solve the simultaneous congruences:
(i) $x \equiv 2 \bmod 5$ and $3x \equiv 1 \bmod 8$;
(ii) $3x \equiv 2 \bmod 5$ and $2x \equiv 1 \bmod 3$.

1.41 (i) Show that $(a+b)^n \equiv a^n + b^n \bmod 2$ for all a and b and for all $n \geq 1$.
Hint. Consider the parity of a and of b.
(ii) Show that $(a+b)^2 \not\equiv a^2 + b^2 \bmod 3$.

1.42 On a desert island, five men and a monkey gather coconuts all day, then sleep. The first man awakens and decides to take his share. He divides the coconuts into five equal shares, with one coconut left over. He gives the extra one to the monkey, hides his share, and goes to sleep. Later, the second man awakens and takes his fifth from the remaining pile; he, too, finds one extra and gives it to the monkey. Each of the remaining three men does likewise in turn. Find the minimum number of coconuts originally present.
Hint. Try -4 coconuts.

1.2 Roots of Unity

Let us now say a bit about the complex numbers $\mathbb{C}$. We define a complex number $z = a + ib$ to be the point (a, b) in the plane; a is called the ***real part*** of z and b is called its ***imaginary part***. The ***modulus*** $|z|$ of $z = a + ib = (a, b)$ is the distance from z to the origin:

$$|z| = \sqrt{a^2 + b^2}.$$

Proposition 1.29 (Polar Decomposition). *Every complex number z has a factorization*

$$z = r(\cos\theta + i \sin\theta),$$

where $r = |z| \geq 0$ *and* $0 \leq \theta < 2\pi$.

Proof. If $z = 0$, then $|z| = 0$, and any choice of θ works. If $z = a + ib \neq 0$, then $|z| \neq 0$, and $z/|z| = (a/|z|, b/|z|)$ has modulus 1, because

$$(a/|z|)^2 + (b/|z|)^2 = (a^2 + b^2)/|z|^2 = 1.$$

Therefore, there is an angle θ (see Figure 1.4 on page 16) with $z/|z| = \cos\theta + i \sin\theta$, and so $z = |z|(\cos\theta + i \sin\theta) = r(\cos\theta + i \sin\theta)$. •

It follows that every complex number z of modulus 1 is a point on the unit circle, and so it has coordinates $(\cos\theta, \sin\theta)$ (θ is the angle from the x-axis to the line joining the origin to (a, b), because $\cos\theta = a/1$ and $\sin\theta = b/1$).

If $z = a + ib = r(\cos\theta + i \sin\theta)$, then (r, θ) are the ***polar coordinates*** of z; this is the reason why Proposition 1.29 is called the polar decomposition of z.

The trigonometric addition formulas for $\cos(\theta + \psi)$ and $\sin(\theta + \psi)$ have a lovely translation into the language of complex numbers.

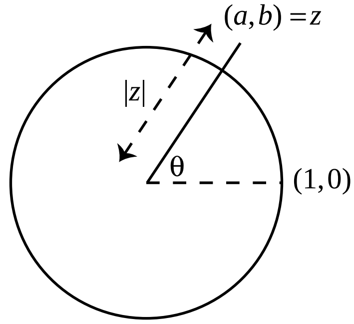


Figure 1.4

Proposition 1.30 (Addition Theorem). *If*

$$z = \cos\theta + i\sin\theta \quad \textit{and} \quad w = \cos\psi + i\sin\psi,$$

then

$$zw = \cos(\theta + \psi) + i\sin(\theta + \psi).$$

Proof.

$$\begin{aligned} zw &= (\cos\theta + i\sin\theta)(\cos\psi + i\sin\psi) \\ &= (\cos\theta\cos\psi - \sin\theta\sin\psi) + i(\sin\theta\cos\psi + \cos\theta\sin\psi). \end{aligned}$$

The trigonometric addition formulas show that

$$zw = \cos(\theta + \psi) + i\sin(\theta + \psi). \quad \bullet$$

The addition theorem gives a geometric interpretation of complex multiplication.

Corollary 1.31. *If z and w are complex numbers with polar coordinates (r, θ) and (s, ψ), respectively, then the polar coordinates of zw are*[2]

$$(rs, \theta + \psi),$$

and so

$$|zw| = |z|\,|w|.$$

Proof. If the polar decompositions of z and w are $z = r(\cos\theta + i\sin\theta)$ and $w = s(\cos\psi + i\sin\psi)$, respectively, then

$$zw = rs[\cos(\theta + \psi) + i\sin(\theta + \psi)]. \quad \bullet$$

[2]This formula is correct if $\theta + \psi \leq 2\pi$; otherwise, the angle should be $\theta + \psi - 2\pi$.

In particular, if $|z| = 1 = |w|$, then $|zw| = 1$; that is, the product of two complex numbers on the unit circle also lies on the unit circle.

In 1707, A. De Moivre (1667–1754) proved the following elegant result.

Theorem 1.32 (*De Moivre*). *For every real number x and every positive integer n,*

$$\cos(nx) + i\sin(nx) = (\cos x + i\sin x)^n.$$

Proof. We prove De Moivre's theorem by induction on $n \geq 1$. The base step $n = 1$ is obviously true. For the inductive step,

$$\begin{aligned}(\cos x + i\sin x)^{n+1} &= (\cos x + i\sin x)^n(\cos x + i\sin x)\\ &= (\cos(nx) + i\sin(nx))(\cos x + i\sin x)\\ &\qquad\qquad \text{(inductive hypothesis)}\\ &= \cos(nx + x) + i\sin(nx + x)\\ &\qquad\qquad \text{(addition formula)}\\ &= \cos([n+1]x) + i\sin([n+1]x). \quad \bullet\end{aligned}$$

Corollary 1.33.

(i) $$\cos(2x) = \cos^2 x - \sin^2 x = 2\cos^2 x - 1$$
$$\sin(2x) = 2\sin x\cos x.$$

(ii) $$\cos(3x) = \cos^3 x - 3\cos x\sin^2 x = 4\cos^3 x - 3\cos x$$
$$\sin(3x) = 3\cos^2 x\sin x - \sin^3 x = 3\sin x - 4\sin^3 x.$$

Proof. (i) $$\begin{aligned}\cos(2x) + i\sin(2x) &= (\cos x + i\sin x)^2\\ &= \cos^2 x + 2i\sin x\cos x + i^2\sin^2 x\\ &= \cos^2 x - \sin^2 x + i(2\sin x\cos x).\end{aligned}$$

Equating real and imaginary parts gives both double angle formulas.

(ii) De Moivre's theorem gives

$$\begin{aligned}\cos(3x) + i\sin(3x) &= (\cos x + i\sin x)^3\\ &= \cos^3 x + 3i\cos^2 x\sin x + 3i^2\cos x\sin^2 x + i^3\sin^3 x\\ &= \cos^3 x - 3\cos x\sin^2 x + i(3\cos^2 x\sin x - \sin^3 x).\end{aligned}$$

Equality of the real parts gives $\cos(3x) = \cos^3 x - 3\cos x\sin^2 x$; the second formula for $\cos(3x)$ follows by replacing $\sin^2 x$ by $1 - \cos^2 x$. Equality of the imaginary parts gives $\sin(3x) = 3\cos^2 x\sin x - \sin^3 x = 3\sin x - 4\sin^3 x$; the second formula arises by replacing $\cos^2 x$ by $1 - \sin^2 x$. $\bullet$

Corollary 1.33 can be generalized. If $f_2(x) = 2x^2 - 1$, then

$$\cos(2x) = 2\cos^2 x - 1 = f_2(\cos x),$$

and if $f_3(x) = 4x^3 - 3x$, then

$$\cos(3x) = 4\cos^3 x - 3\cos x = f_3(\cos x).$$

Proposition 1.34. *For all $n \geq 1$, there is a polynomial $f_n(x)$ having all coefficients integers such that*

$$\cos(nx) = f_n(\cos x).$$

Proof. By De Moivre's theorem,

$$\begin{aligned}\cos(nx) + i\sin(nx) &= (\cos x + i\sin x)^n \\ &= \sum_{r=0}^{n} \binom{n}{r} \cos^{n-r} x\, i^r \sin^r x.\end{aligned}$$

The real part of the left side, $\cos(nx)$, must be equal to the real part of the right side. Now i^r is real if and only if r is even, and so

$$\cos(nx) = \sum_{r \text{ even}} \binom{n}{r} \cos^{n-r} x\, i^r \sin^r x.$$

If $r = 2k$, then $i^r = i^{2k} = (-1)^k$, and

$$\cos(nx) = \sum_{k=0}^{\left[\frac{n}{2}\right]} (-1)^k \binom{n}{2k} \cos^{n-2k} x \sin^{2k} x,$$

where $\left[\frac{n}{2}\right]$ is the largest integer $\leq \frac{n}{2}$. But $\sin^{2k} x = (\sin^2 x)^k = (1 - \cos^2 x)^k$, which is a polynomial in $\cos x$. This completes the proof. •

It is not difficult to show, by induction on $n \geq 2$, that $f_n(x)$ begins with $2^{n-1}x^n$. A sine version of Proposition 1.34 can be found in Exercise 1.49 on page 25.

Euler's Theorem. *For all real numbers x,*

$$e^{ix} = \cos x + i\sin x.$$

The basic idea of the proof, aside from matters of convergence, is to examine the real and imaginary parts of the power series expansion of e^{ix}. Using the fact that the powers of i repeat in cycles of length 4: $1, i, -1, -i, 1 \ldots$, we have

$$\begin{aligned}e^{ix} &= 1 + ix + \frac{(ix)^2}{2!} + \frac{(ix)^3}{3!} + \cdots \\ &= \left[1 - \frac{x^2}{2!} + \frac{x^4}{4!} + \cdots\right] + i\left[x - \frac{x^3}{3!} + \frac{x^5}{5!} + \cdots\right] \\ &= \cos x + i\sin x.\end{aligned}$$

It is said that Euler was especially pleased with the equation

$$e^{\pi i} = -1;$$

indeed, this formula is inscribed on his tombstone.

As a consequence of Euler's theorem, the polar decomposition can be rewritten in exponential form: Every complex number z has a factorization

$$z = re^{i\theta},$$

where $r \geq 0$ and $0 \leq \theta < 2\pi$. The addition theorem and De Moivre's theorem can be restated in complex exponential form. The first becomes

$$e^{ix}e^{iy} = e^{i(x+y)};$$

the second becomes

$$(e^{ix})^n = e^{inx}.$$

Definition. If $n \geq 1$ is an integer, then an ***nth root of unity*** is a complex number ζ with $\zeta^n = 1$.

The geometric interpretation of complex multiplication is particularly interesting when z and w lie on the unit circle, so that $|z| = 1 = |w|$. Given a positive integer n, let $\theta = 2\pi/n$ and let $\zeta = e^{i\theta}$. The polar coordinates of ζ are $(1, \theta)$, the polar coordinates of ζ^2 are $(1, 2\theta)$, the polar coordinates of ζ^3 are $(1, 3\theta)$,$\ldots$, the polar coordinates of ζ^{n-1} are $(1, (n-1)\theta)$, and the polar coordinates of $\zeta^n = 1$ are $(1, n\theta) = (1, 0)$. Thus, the nth roots of unity are equally spaced around the unit circle. Figure 1.5 shows the 8th roots of unity (here, $\theta = 2\pi/8 = \pi/4$).

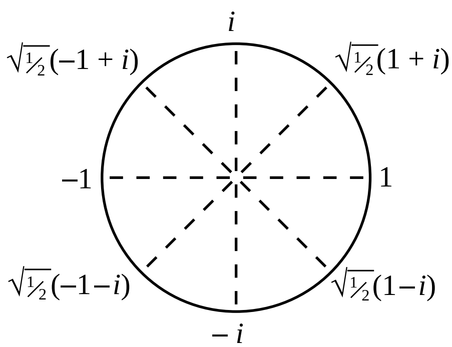


Figure 1.5: 8th Roots of Unity

Corollary 1.35. *Every nth root of unity is equal to*

$$e^{2\pi ik/n} = \cos\left(\tfrac{2\pi k}{n}\right) + i\sin\left(\tfrac{2\pi k}{n}\right),$$

for some $k = 0, 1, 2, \ldots, n-1$, and hence it has modulus 1.

Proof. Note that $e^{2\pi i} = \cos 2\pi + i \sin 2\pi = 1$. By De Moivre's theorem, if $\zeta = e^{2\pi i/n} = \cos(2\pi/n) + i\sin(2\pi/n)$, then

$$\zeta^n = (e^{2\pi i/n})^n = e^{2\pi i} = 1,$$

so that ζ is an nth root of unity. Since $\zeta^n = 1$, it follows that $(\zeta^k)^n = (\zeta^n)^k = 1^k = 1$ for all $k = 0, 1, 2, \ldots, n-1$, so that $\zeta^k = e^{2\pi ik/n}$ is also an nth root of unity. We have exhibited n distinct nth roots of unity; there can be no others, for it will be proved in Chapter 3 that a polynomial of degree n with rational coefficients (e.g., $x^n - 1$) has at most n complex roots. •

Just as there are two square roots of a number a, namely, $\sqrt{a}$ and $-\sqrt{a}$, there are n different nth roots of a, namely, $e^{2\pi ik/n}\sqrt[n]{a}$ for $k = 0, 1, \ldots, n-1$.

Every nth root of unity is, of course, a root of the polynomial $x^n - 1$. Therefore,

$$x^n - 1 = \prod_{\zeta^n = 1} (x - \zeta).$$

If ζ is an nth root of unity, and if n is the smallest positive integer for which $\zeta^n = 1$, we say that ζ is a ***primitive nth root of unity***. Thus, i is an 8th root of unity, but it is not a primitive 8th root of unity; however, i is a primitive 4th root of unity.

Lemma 1.36. *If an nth root of unity ζ is a primitive dth root of unity, then d must be a divisor of n.*

Proof. The division algorithm gives $n = qd + r$, where q are r are integers and the remainder r satisfies $0 \le r < d$. But

$$1 = \zeta^n = \zeta^{qd+r} = \zeta^{qd}\zeta^r = \zeta^r,$$

because $\zeta^{qd} = (\zeta^d)^q = 1$. If $r \neq 0$, we contradict d being the smallest exponent for which $\zeta^d = 1$. Hence, $n = qd$, as claimed. •

Definition. If d is a positive integer, then the dth ***cyclotomic***[3] ***polynomial*** is defined by

$$\Phi_d(x) = \prod (x - \zeta),$$

where ζ ranges over all the *primitive* dth roots of unity.

The following result is almost obvious.

[3] The roots of $x^n - 1$ are the nth roots of unity: $1, \zeta, \zeta^2, \ldots, \zeta^{n-1}$, where $\zeta = e^{2\pi i/n} = \cos(2\pi/n) + i\sin(2\pi/n)$. Now these roots divide the unit circle $\{\zeta \in \mathbb{C} : |z| = 1\}$ into n equal arcs (see Figure 1.5 on page 19). This explains the term *cyclotomic*, for its Greek origin means "circle splitting."

Proposition 1.37. *For every integer $n \geq 1$,*

$$x^n - 1 = \prod_{d|n} \Phi_d(x),$$

where d ranges over all the divisors d of n [in particular, $\Phi_1(x)$ and $\Phi_n(x)$ occur].

Proof. In light of Corollary 1.35, the proposition follows by collecting, for each divisor d of n, all terms in the equation $x^n - 1 = \prod(x - \zeta)$ with ζ a primitive dth root of unity. •

For example, if p is a prime, then $x^p - 1 = \Phi_1(x)\Phi_p(x)$. Since $\Phi_1(x) = x - 1$, it follows that

$$\Phi_p(x) = x^{p-1} + x^{p-2} + \cdots + x + 1.$$

Definition. Define the ***Euler ϕ-function*** as the degree of the nth cyclotomic polynomial:

$$\phi(n) = \deg(\Phi_n(x)).$$

We now give another description of the Euler ϕ-function that does not depend on roots of unity.

Proposition 1.38. *If $n \geq 1$ is an integer, then $\phi(n)$ is the number of integers k with $1 \leq k \leq n$ and $(k, n) = 1$.*

Proof. It suffices to prove that $e^{2\pi ik/n}$ is a primitive nth root of unity if and only if k and n are relatively prime.

If k and n are not relatively prime, then $n = dr$ and $k = ds$, where d, r, and s are integers, and $d > 1$; it follows that $r < n$. Hence, $\frac{k}{n} = \frac{ds}{dr} = \frac{s}{r}$, so that $(e^{2\pi ik/n})^r = (e^{2\pi is/r})^r = 1$, and hence $e^{2\pi ik/n}$ is not a primitive nth root of unity.

Conversely, suppose that $\zeta = e^{2\pi ik/n}$ is not a primitive nth root of unity. Lemma 1.36 says that ζ must be a dth root of unity for some divisor d of n with $d < n$; that is, there is $1 \leq m \leq d$ with

$$\zeta = e^{2\pi ik/n} = e^{2\pi im/d} = e^{2\pi imr/dr} = e^{2\pi imr/n}.$$

Since both k and mr are in the range between 1 and n, it follows that $k = mr$ (if $0 \leq x, y < 1$ and $e^{2\pi ix} = e^{2\pi iy}$, then $x = y$); that is, r is a divisor of k and of n, and so k and n are not relatively prime. •

Corollary 1.39. *For every integer $n \geq 1$, we have*

$$n = \sum_{d|n} \phi(d).$$

Proof. Note that $\phi(n)$ is the degree of $\Phi_n(x)$, and use the fact that the degree of a product of polynomials is the sum of the degrees of the factors. •

Recall that the ***leading coefficient*** of a polynomial $f(x)$ is the coefficient of the highest power of x occurring in $f(x)$; we say that a polynomial $f(x)$ is ***monic*** if its leading coefficient is 1.

Proposition 1.40. *For every positive integer n, the cyclotomic polynomial $\Phi_n(x)$ is a monic polynomial all of whose coefficients are integers.*

Proof. The proof is by induction on $n \geq 1$. The base step holds, for $\Phi_1(x) = x - 1$. For the inductive step, we assume that $\Phi_d(x)$ is a monic polynomial with integer coefficients. From the equation $x^n - 1 = \prod_d \Phi_d(x)$, we have

$$x^n - 1 = \Phi_n(x) f(x),$$

where $f(x)$ is the product of all $\Phi_d(x)$, where $d < n$ and d is a divisor of n. By the inductive hypothesis, $f(x)$ is a monic polynomial with integer coefficients. Because $f(x)$ is monic, long division (i.e., the division algorithm for polynomials) shows that all the coefficients of $\Phi_n(x) = (x^n - 1)/f(x)$ are also integers,[4] as desired. •

The following corollary will be used in Chapter 8 to prove a theorem of Wedderburn.

Corollary 1.41. *If q is a positive integer, and if d is a divisor of an integer n with $d < n$, then $\Phi_n(q)$ is a divisor of both $q^n - 1$ and $(q^n - 1)/(q^d - 1)$.*

Proof. We have already seen that $x^n - 1 = \Phi_n(x) f(x)$, where $f(x)$ is a monic polynomial with integer coefficients. Setting $x = q$ gives an equation in integers: $q^n - 1 = \Phi_n(q) f(q)$; that is, $\Phi_n(q)$ is a divisor of $q^n - 1$.

If d is a divisor of n and $d < n$, consider the equation $x^d - 1 = \prod(x - \zeta)$, where ζ ranges over the dth roots of unity. Notice that each such ζ is an nth root of unity, because d is a divisor of n. Since $d < n$, collecting terms in the equation $x^n - 1 = \prod(x - \zeta)$ gives

$$x^n - 1 = \Phi_n(x)(x^d - 1)g(x),$$

where $g(x)$ is the product of all the cyclotomic polynomials $\Phi_\delta(x)$ for all divisors δ of n with $\delta < n$ and with δ not a divisor of d. It follows from the proposition that $g(x)$ is a monic polynomial with integer coefficients. Therefore, $g(q) \in \mathbb{Z}$ and

$$\frac{x^n - 1}{x^d - 1} = \Phi_n(x)g(x)$$

gives the result. •

Here is the simplest way to find the reciprocal of a complex number. If $z = a + ib \in \mathbb{C}$, where $a, b \in \mathbb{R}$, define its ***complex conjugate*** $\overline{z} = a - ib$. Note that $z\overline{z} = a^2 + b^2 = |z|^2$, so that $z \neq 0$ if and only if $z\overline{z} \neq 0$. If $z \neq 0$, then

$$z^{-1} = 1/z = \overline{z}/z\overline{z} = (a/z\overline{z}) - i(b/z\overline{z});$$

that is,

$$\frac{1}{a + ib} = \left(\frac{a}{a^2 + b^2}\right) - i\left(\frac{b}{a^2 + b^2}\right).$$

[4] If this is not clear, look at the proof of the division algorithm on page 131.

If $|z| = 1$, then $z^{-1} = \overline{z}$. In particular, if z is a root of unity, then its reciprocal is its complex conjugate.

Complex conjugation satisfies the following identities:

$$\begin{aligned} \overline{z+w} &= \overline{z} + \overline{w}; \\ \overline{zw} &= \overline{z}\,\overline{w}; \\ \overline{\overline{z}} &= z; \\ \overline{z} &= z \quad \text{if and only if } z \text{ is real.} \end{aligned}$$

We are regarding complex numbers as points in the plane and, as in vector calculus, a point z is identified with the vector represented by the arrow $\overrightarrow{Oz}$ from the origin O to z. Let us define the ***dot product*** of $z = a + ib$ and $w = c + id$ to be

$$z \cdot w = ac + bd.$$

Thus, $z \cdot w = |z||w| \cos\theta$, where θ is the angle between $\overrightarrow{Oz}$ and $\overrightarrow{Ow}$ [since $\cos\theta = \cos(2\pi - \theta)$, it makes no difference whether θ is measured from $\overrightarrow{Oz}$ to $\overrightarrow{Ow}$ or from $\overrightarrow{Ow}$ to $\overrightarrow{Oz}$]. Note that

$$z \cdot z = |z|^2 = z\overline{z}.$$

It is clear that $z \cdot w = w \cdot z$, and it is easy to check that

$$z \cdot (w + w') = z \cdot w + z \cdot w'$$

for all complex numbers z, w, and w'.

The following result will be used in Chapter 8 to prove a theorem of Burnside.

Proposition 1.42. *If $\varepsilon_1, \ldots, \varepsilon_n$ are roots of unity, where $n \geq 2$, then*

$$\Big|\sum_{j=1}^{n} \varepsilon_j\Big| \leq \sum_{j=1}^{n} \big|\varepsilon_j\big| = n.$$

Moreover, there is equality if and only if all the ε_j are equal.

Proof. The proof of the inequality is by induction on $n \geq 2$. The base step follows from the ***triangle inequality***: for all complex numbers u and v,

$$|u + v| \leq |u| + |v|.$$

The proof of the inductive step is routine, for roots of unity have modulus 1.

Suppose now that all the ε_j are equal, say $\varepsilon_j = \varepsilon$ for all j, then it is clear that there is equality $\big|\sum_{j=1}^{n} \varepsilon_j\big| = |n\varepsilon| = n|\varepsilon| = n$. The proof of the converse is by induction on $n \geq 2$. For the base step, suppose that $|\varepsilon_1 + \varepsilon_2| = 2$. Using the dot product, we have

$$\begin{aligned} 4 &= |\varepsilon_1 + \varepsilon_2|^2 \\ &= (\varepsilon_1 + \varepsilon_2) \cdot (\varepsilon_1 + \varepsilon_2) \\ &= |\varepsilon_1|^2 + 2\varepsilon_1 \cdot \varepsilon_2 + |\varepsilon_2|^2 \\ &= 2 + 2\varepsilon_1 \cdot \varepsilon_2. \end{aligned}$$

Hence, $2 = 1 + \varepsilon_1 \cdot \varepsilon_2$, so that

$$\begin{aligned} 1 &= \varepsilon_1 \cdot \varepsilon_2 \\ &= |\varepsilon_1||\varepsilon_2| \cos\theta \\ &= \cos\theta, \end{aligned}$$

where θ is the angle between $\overrightarrow{O\varepsilon_1}$ and $\overrightarrow{O\varepsilon_2}$ (for $|\varepsilon_1| = 1 = |\varepsilon_2|$). Therefore, $\theta = 0$ or $\theta = \pi$, so that $\varepsilon_2 = \pm\varepsilon_1$. Since $\varepsilon_2 = -\varepsilon_1$ gives $|\varepsilon_1 + \varepsilon_2| = 0$, we must have $\varepsilon_2 = \varepsilon_1$.

For the inductive step, suppose that $\left|\sum_{j=1}^{n+1} \varepsilon_j\right| = n + 1$. If $\left|\sum_{j=1}^{n} \varepsilon_j\right| < n$, then the triangle inequality gives

$$\left|\Big(\sum_{j=1}^{n} \varepsilon_j\Big) + \varepsilon_{n+1}\right| \le \left|\sum_{j=1}^{n} \varepsilon_j\right| + 1 < n + 1,$$

contrary to hypothesis. Therefore, $\left|\sum_{j=1}^{n} \varepsilon_j\right| = n$, and so the inductive hypothesis gives $\varepsilon_1, \ldots, \varepsilon_n$ all equal, say, to ω. Hence, $\sum_{j=1}^{n} \varepsilon_j = n\omega$, and so

$$|n\omega + \varepsilon_{n+1}| = n + 1.$$

The argument concludes as that of the base step:

$$\begin{aligned} (n+1)^2 &= (n\omega + \varepsilon_{n+1}) \cdot (n\omega + \varepsilon_{n+1}) \\ &= n^2 + 2n\omega \cdot \varepsilon_{n+1} + 1, \end{aligned}$$

so that $1 = \omega \cdot \varepsilon_{n+1} = |\omega||\varepsilon_{n+1}| \cos\theta$, where θ is the angle between $\overrightarrow{O\omega}$ and $\overrightarrow{O\varepsilon_{n+1}}$. Hence, $\omega = \pm\varepsilon_{n+1}$, and $\omega = \varepsilon_{n+1}$. •

EXERCISES

1.43 Evaluate $(\cos 3^\circ + i \sin 3^\circ)^{40}$.

1.44 (i) Find $(3 + 4i)/(2 - i)$.

(ii) If $z = re^{i\theta}$, prove that $z^{-1} = r^{-1}e^{-i\theta}$.

(iii) Find the values of $\sqrt{i}$.

(iv) Prove that $e^{i\theta/n}$ is an nth root of $e^{i\theta}$.

1.45 Find $\Phi_6(x)$.

1.46 If α is a number for which $\cos(\pi\alpha) = \frac{1}{3}$ (where the angle $\pi\alpha$ is in radians), prove that α is irrational.

Hint. If $\alpha = \frac{m}{n}$, evaluate $\cos n\pi\alpha + i \sin n\pi\alpha$ using De Moivre's theorem.

1.47 Let $f(x) = a_0 + a_1x + a_2x^2 + \cdots + a_nx^n$ be a polynomial with all of its coefficients real numbers. Prove that if z is a root of $f(x)$, then $\overline{z}$ is also a root of $f(x)$.

1.48 (i) Prove that the quadratic formula holds for quadratic polynomials with complex coefficients.

(ii) Find the roots of $x^2 + (2 + i)x + 2i$. Why aren't the roots complex conjugates of one another?

1.49 Prove that for every *odd* integer $n \geq 1$, there is a polynomial $g_n(x)$ with integer coefficients, such that

$$\sin nx = g_n(\sin x).$$

1.50 Every Pythagorean triple (a, b, c) determines a right triangle having legs a and b and hypotenuse[5] c. Call two Pythagorean triples (a, b, c) and (a', b', c') ***similar*** if the right triangles they determine are similar triangles; that is, if corresponding sides are proportional.

(i) Prove that the following statements are equivalent for Pythagorean triples (a, b, c) and (a', b', c').

(1) (a, b, c) and (a', b', c') are similar.

(2) There are positive integers m and ℓ with $(ma, mb, mc) = (\ell a', \ell b', \ell c')$

(3) $\frac{a}{c} + i\frac{b}{c} = \frac{a'}{c'} + i\frac{b'}{c'}$.

(ii) Prove that every Pythagorean triple is similar to a primitive Pythagorean triple.

1.51 (i) Call a complex number of modulus 1 ***rational*** if both its real and imaginary parts are rational numbers. If $\frac{a}{c} + i\frac{b}{c}$ is a rational complex number with both a and b nonzero, prove that $(|a|, |b|, |c|)$ is a Pythagorean triple.

(ii) Prove that the product of two rational complex numbers is also a rational complex number, and use this fact to define a product of two Pythagorean triples (up to similarity). What is the product of $(3, 4, 5)$ with itself?

(iii) Show that the square of a Pythagorean triple (a, b, c) is $(a^2 - b^2, 2ab, a^2 + b^2)$.

1.3 SOME SET THEORY

Functions are ubiquitous in algebra, as in all of mathematics, and we discuss them now.

A set X is a collection of elements (numbers, points, herring, etc.); we write

$$x \in X$$

to denote x belonging to X. Two sets X and Y are defined to be ***equal***, denoted by

$$X = Y,$$

if they are comprised of exactly the same elements; for every element x, we have $x \in X$ if and only if $x \in Y$.

A ***subset*** of a set X is a set S each of whose elements also belongs to X: If $s \in S$, then $s \in X$. We denote S being a subset of X by

$$S \subseteq X;$$

[5]*Hypotenuse* comes from the Greek word meaning "to stretch."

synonyms are “S is ***contained*** in X” and “S is ***included*** in X.” Note that $X \subseteq X$ is always true; we say that a subset S of X is a ***proper subset*** of X, denoted by $S \subsetneq X$, if $S \subseteq X$ and $S \neq X$. It follows from the definitions that two sets X and Y are equal if and only if each is a subset of the other:

$$X = Y \qquad \text{if and only if} \qquad X \subseteq Y \text{ and } Y \subseteq X.$$

Because of this remark, many proofs showing that two sets are equal break into two parts, each half showing that one of the sets is a subset of the other. For example, let

$$X = \{a \in \mathbb{R} : a \geq 0\} \qquad \text{and} \qquad Y = \{r^2 : r \in \mathbb{R}\}.$$

If $a \in X$, then $a \geq 0$ and $a = r^2$, where $r = \sqrt{a}$; hence, $a \in Y$ and $X \subseteq Y$. For the reverse inclusion, choose $r^2 \in Y$. If $r \geq 0$, then $r^2 \geq 0$; if $r < 0$, then $r = -s$, where $s > 0$, and $r^2 = (-1)^2 s^2 = s^2 \geq 0$. In either case, $r^2 \geq 0$ and $r^2 \in X$. Therefore, $Y \subseteq X$, and so $X = Y$.

Calculus books define a function $f(x)$ as a “rule” that assigns, to each number a, exactly one number, namely, $f(a)$. This definition is certainly in the right spirit, but it has a defect: What is a rule? To ask this question another way, when are two rules the same? For example, consider the functions

$$f(x) = (x+1)^2 \qquad \text{and} \qquad g(x) = x^2 + 2x + 1.$$

Is $f(x) = g(x)$? The evaluation procedures are certainly different: for example, $f(6) = (6+1)^2 = 7^2$, while $g(6) = 6^2 + 2 \cdot 6 + 1 = 36 + 12 + 1$. Since the term *rule* has not been defined, it is ambiguous, and our question has no answer. Surely the calculus description is inadequate if we cannot decide whether these two functions are the same.

The graph of a function is a concrete thing [for example, the graph of $f(x) = x^2$ is a parabola], and the upcoming formal definition of a function amounts to saying that a function *is* its graph. The informal calculus definition of a function as a rule remains, but we will have avoided the problem of saying what a rule is. In order to give the definition, we first need an analog of the plane [for we will want to use functions $f(x)$ whose argument x does not vary over numbers].

Definition. If X and Y are (not necessarily distinct) sets, then their ***cartesian product*** $X \times Y$ is the set of all ordered pairs (x, y), where $x \in X$ and $y \in Y$.

The plane is $\mathbb{R} \times \mathbb{R}$.

The only thing we need to know about ordered pairs is that

$$(x, y) = (x', y') \quad \text{if and only if} \quad x = x' \text{ and } y = y'$$

(see Exercise 1.62 on page 37).

Observe that if X and Y are finite sets, say, $|X| = m$ and $|Y| = n$ (we denote the number of elements in a finite set X by $|X|$), then $|X \times Y| = mn$.

Definition. Let X and Y be (not necessarily distinct) sets. A ***function*** f from X to Y, denoted by[6]

$$f\colon X \to Y,$$

is a subset $f \subseteq X \times Y$ such that, for each $a \in X$, there is a unique $b \in Y$ with $(a, b) \in f$.

For each $a \in X$, the unique element $b \in Y$ for which $(a, b) \in f$ is called the ***value*** of f at a, and b is denoted by $f(a)$. Thus, f consists of all those points in $X \times Y$ of the form $(a, f(a))$. When $f : \mathbb{R} \to \mathbb{R}$, then f *is* the graph of $f(x)$.

Call X the ***domain*** of f, call Y the ***target*** (or *codomain*) of f, and define the ***image*** (or *range*) of f, denoted by $\operatorname{im} f$, to be the subset of Y consisting of all the values of f.

Definition. Two functions $f\colon X \to Y$ and $g\colon X' \to Y'$ are ***equal*** if $X = X'$, $Y = Y'$, and the subsets $f \subseteq X \times Y$ and $g \subseteq X' \times Y'$ are equal.

For example, if X is a set, then the ***identity function*** $1_X\colon X \to X$ is defined by $1_X(x) = x$ for all $x \in X$; if $X = \mathbb{R}$, then $1_{\mathbb{R}}$ is the line with slope 1 that passes through the origin. If $f\colon X \to Y$ is a function, and if S is a subset of X, then the ***restriction*** of f to S is the function $f|S\colon S \to Y$ defined by $(f|S)(s) = f(s)$ for all $s \in S$. If S is a subset of a set X, define the ***inclusion*** $i\colon S \to X$ to be the function defined by $i(s) = s$ for all $s \in S$. If S is a proper subset of X, then the inclusion i is not the identity function 1_S because its target is X, not S; it is not the identity function 1_X because its domain is S, not X.

A function $f\colon X \to Y$ has three ingredients: its domain X, its target Y, and its graph, and we are saying that two functions are equal if and only if they have the same domains, the same targets, and the same graphs.

It is plain that the domain and the graph are essential parts of a function. Why should we care about the target of a function when its image is more important?

As a practical matter, when first defining a function, we usually do not know its image. For example, we say that $f\colon \mathbb{R} \to \mathbb{R}$, given by $f(x) = x^2 + 3x - 8$, is a real-valued function, and we then analyze f to find its image. But if targets have to be images, then we could not even write down $f\colon X \to Y$ without having first found the image of f (and finding the precise image is often very difficult, if not impossible); thus, targets are convenient to use.

In linear algebra, we consider a vector space V and its *dual space* $V^* = \{$all linear functionals on $V\}$ (which is also a vector space). Moreover, every linear transformation $T\colon V \to W$ defines a linear transformation

$$T^*\colon W^* \to V^*,$$

and the domain of T^*, being W^*, is determined by the target W of T. (In fact, if a matrix for T is A, then a matrix for T^* is A^t, the transpose of A.) Thus, changing the target of T changes the domain of T^*, and so T^* is changed in an essential way.

---

[6]From now on, we denote a function by f instead of by $f(x)$. The notation $f(x)$ is reserved for the value of f at x; there are a few exceptions: We will continue to write $\sin x$, e^x, and x^2, for example.

Proposition 1.43. *Let $f: X \to Y$ and $g: X \to Y$ be functions. Then $f = g$ if and only if $f(a) = g(a)$ for every $a \in X$.*

Remark. This proposition resolves the problem raised by the ambiguous term *rule*. If f, $g: \mathbb{R} \to \mathbb{R}$ are given by $f(x) = (x+1)^2$ and $g(x) = x^2 + 2x + 1$, then $f = g$ because $f(a) = g(a)$ for every number a. ◀

Proof. Assume that $f = g$. Functions are subsets of $X \times Y$, and so $f = g$ means that each of f and g is a subset of the other (informally, we are saying that f and g have the same graph). If $a \in X$, then $(a, f(a)) \in f = g$, and so $(a, f(a)) \in g$. But there is only one ordered pair in g with first coordinate a, namely, $(a, g(a))$ (because the definition of function says that g gives a unique value to a). Therefore, $(a, f(a)) = (a, g(a))$, and equality of ordered pairs gives $f(a) = g(a)$, as desired.

Conversely, assume that $f(a) = g(a)$ for every $a \in X$. To see that $f = g$, it suffices to show that $f \subseteq g$ and $g \subseteq f$. Each element of f has the form $(a, f(a))$. Since $f(a) = g(a)$, we have $(a, f(a)) = (a, g(a))$, and hence $(a, f(a)) \in g$. Therefore, $f \subseteq g$. The reverse inclusion $g \subseteq f$ is proved similarly. •

We continue to regard a function f as a rule sending $x \in X$ to $f(x) \in Y$, but the precise definition is now available whenever we need it, as in the proof of Proposition 1.43. However, to reinforce our wanting to regard functions $f: X \to Y$ as dynamic things sending points in X to points in Y, we often write

$$f: x \mapsto y$$

instead of $f(x) = y$. For example, we may write $f: x \mapsto x^2$ instead of $f(x) = x^2$, and we may describe the identity function by $x \mapsto x$ for all x.

Instead of saying that values of a function f are unique, we usually say that f is ***well-defined*** (or *single-valued*). Does the formula $g(a/b) = ab$ define a function $g: \mathbb{Q} \to \mathbb{Q}$? There are many ways to write a fraction; since $\frac{1}{2} = \frac{3}{6}$, we see that $g(\frac{1}{2}) = 1 \cdot 2 \neq 3 \cdot 6 = g\left(\frac{3}{6}\right)$, and so g is not a function because it is not well-defined. Had we said that the formula $g(a/b) = ab$ holds whenever a/b is in lowest terms, then g would be a function.

The formula $f(a/b) = 3a/b$ does define a function $f: \mathbb{Q} \to \mathbb{Q}$, for it is well-defined: If $a/b = a'/b'$, we show that

$$f(a/b) = 3a/b = 3a'/b' = f(a'/b'):$$

$a/b = a'/b'$ gives $ab' = a'b$, so that $3ab' = 3a'b$ and $3a/b = 3a'/b'$. Thus, f is a bona fide function; that is, f is well-defined.

Example 1.44.

Our definitions allow us to treat a degenerate case. If X is a set, what are the functions $X \to \varnothing$? Note first that an element of $X \times \varnothing$ is an ordered pair (x, y) with $x \in X$ and $y \in \varnothing$; since there is no $y \in \varnothing$, there are no such ordered pairs, and so $X \times \varnothing = \varnothing$. Now

a function $X \to \varnothing$ is a subset of $X \times \varnothing$ of a certain type; but $X \times \varnothing = \varnothing$, so there is only one subset, namely $\varnothing$, and hence at most one function, namely, $f = \varnothing$. The definition of function $X \to \varnothing$ says that, for each $x \in X$, there exists a unique $y \in \varnothing$ with $(x, y) \in f$. If $X \neq \varnothing$, then there exists $x \in X$ for which no such y exists (there are no elements y at all in $\varnothing$), and so f is not a function. Thus, if $X \neq \varnothing$, there are no functions from X to $\varnothing$. On the other hand, if $X = \varnothing$, then $f = \varnothing$ is a function. Otherwise, the negation of the statement "f is a function" begins "there exists $x \in \varnothing$, etc." We need not go on; since $\varnothing$ has no elements in it, there is no way to complete the sentence so that it is a true statement. We conclude that $f = \varnothing$ is a function $\varnothing \to \varnothing$, and we declare it to be the identity function $1_\varnothing$. ◀

The special case when the image of a function is the whole target has a name.

Definition. A function $f : X \to Y$ is a ***surjection*** (or is *onto*) if

$$\operatorname{im} f = Y.$$

Thus, f is surjective if, for each $y \in Y$, there is some $x \in X$ (probably depending on y) with $y = f(x)$.

The following definition gives another important property a function may have.

Definition. A function $f : X \to Y$ is an ***injection*** (or is *one-to-one*) if, whenever a and a' are distinct elements of X, then $f(a) \neq f(a')$. Equivalently (the contrapositive states that) f is injective if, for every pair $a, a' \in X$, we have

$$f(a) = f(a') \text{ implies } a = a'.$$

The reader should note that being injective is the converse of being well-defined: f is well-defined if $a = a'$ implies $f(a) = f(a')$; f is injective if $f(a) = f(a')$ implies $a = a'$.

There are other names for these functions. Surjections are often called ***epimorphisms*** and injections are often called ***monomorphisms***. The notation $A \twoheadrightarrow B$ is used to denote a surjection, and the notations $A \hookrightarrow B$ or $A \rightarrowtail B$ are used to denote injections. However, we shall not use this terminology or these notations in this book.

Example 1.45.

Consider the function $f : \mathbb{R} \to \mathbb{R}$, given by $f(x) = 3x - 4$. To see whether f is surjective, take $y \in \mathbb{R}$ and ask whether there is $a \in \mathbb{R}$ with $y = 3a - 4$. We solve to obtain $a = \frac{1}{3}(y + 4)$, and we conclude that f is surjective. Also, the function f is injective, for if $3a - 4 = 3b - 4$, then $a = b$.

As a second example, consider the function $g : \mathbb{R} - \{1\} \to \mathbb{R}$ given by

$$g(x) = \frac{3x - 4}{x - 1}.$$

Now g is an injection, for if $(3a-4)/(a-1) = (3b-4)/(b-1)$, then cross multiplying gives $a = b$. On the other hand, g is not surjective. Given $y \in \mathbb{R}$, is there $a \in \mathbb{R}$ with $y = (3a-4)/(a-1)$? Solving, $a = (4-y)/(3-y)$. This suggests that $y = 3$ is not a value of g, and, indeed, it is not: $3 = (3a-4)/(a-1)$ is not solvable. ◂

Definition. If $f \colon X \to Y$ and $g \colon Y \to Z$ are functions (note that the target of f is equal to the domain of g), then their ***composite***, denoted by $g \circ f$, is the function $X \to Z$ given by

$$g \circ f \colon x \mapsto g(f(x));$$

that is, first evaluate f on x, and then evaluate g on $f(x)$.

The chain rule in calculus is a formula for the derivative $(g \circ f)'$ in terms of g' and f':

$$(g \circ f)' = \big[g' \circ f\big] \cdot f'.$$

For example,

$$(\sin(\ln x))' = \cos(\ln x) \cdot \frac{1}{x}.$$

Given a set X, let

$$\mathcal{F}(X) = \{\text{all functions } X \to X\}.$$

We have just seen that the composite of two functions in $\mathcal{F}(X)$ is always defined; moreover, the composite is again a function in $\mathcal{F}(X)$. We may thus regard $\mathcal{F}(X)$ as being equipped with a kind of multiplication. This multiplication is not ***commutative***; that is, $f \circ g$ and $g \circ f$ need not be equal. For example, if $f(x) = x+1$ and $g(x) = x^2$, then $f \circ g \colon 1 \mapsto 1^2+1 = 2$ while $g \circ f \colon 1 \mapsto (1+1)^2 = 4$; therefore, $f \circ g \neq g \circ f$.

Lemma 1.46.

(i) *Composition is **associative**: If*

$$f \colon X \to Y, \quad g \colon Y \to Z, \quad \textit{and} \quad h \colon Z \to W$$

are functions, then

$$h \circ (g \circ f) = (h \circ g) \circ f.$$

(ii) *If $f \colon X \to Y$, then $1_Y \circ f = f = f \circ 1_X$.*

Sketch of Proof. Use Proposition 1.43. •

Are there "reciprocals" in $\mathcal{F}(X)$; that is, are there any functions f for which there is $g \in \mathcal{F}(X)$ with $f \circ g = 1_X$ and $g \circ f = 1_X$?

Definition. A function $f \colon X \to Y$ is a ***bijection*** (or a *one-to-one correspondence*) if it is both an injection and a surjection.

Definition. A function $f : X \to Y$ has an ***inverse*** if there is a function $g : Y \to X$ with both composites $g \circ f$ and $f \circ g$ being identity functions.

Proposition 1.47.

(i) *If $f : X \to Y$ and $g : Y \to X$ are functions such that $g \circ f = 1_X$, then f is injective and g is surjective.*

(ii) *A function $f : X \to Y$ has an inverse $g : Y \to X$ if and only if f is a bijection.*

Proof. (i) Suppose that $f(x) = f(x')$; apply g to obtain $g(f(x)) = g(f(x'))$; that is, $x = x'$ [because $g(f(x)) = x$], and so f is injective. If $x \in X$, then $x = g(f(x))$, so that $x \in \operatorname{im} g$; hence g is surjective.

(ii) If f has an inverse g, then part (i) shows that f is injective and surjective, for both composites $g \circ f$ and $f \circ g$ are identities.

Assume that f is a bijection. For each $y \in Y$, there is $a \in X$ with $f(a) = y$, since f is surjective, and this element a is unique because f is injective. Defining $g(y) = a$ thus gives a (well-defined) function whose domain is Y, and it is plain that g is the inverse of f; that is, $f(g(y)) = f(a) = y$ for all $y \in Y$ and $g(f(a)) = g(y) = a$ for all $a \in X$. •

Remark. Exercise 1.59 on page 36 shows that if both f and g are injective, then so is their composite $f \circ g$. Similarly, $f \circ g$ is a surjection if both f and g are surjections. It follows that the composite of two bijections is itself a bijection. ◂

Notation. The inverse of a bijection f is denoted by f^{-1} (Exercise 1.54 on page 36 says that a function cannot have two inverses).

Example 1.48.

Here is an example of two functions f and g one of whose composites $g \circ f$ is the identity while the other composite $f \circ g$ is not the identity; thus, f and g are not inverse functions.

If $\mathbb{N} = \{n \in \mathbb{Z} : n \geq 0\}$, define $f, g : \mathbb{N} \to \mathbb{N}$ as follows:

$$f(n) = n + 1;$$
$$g(n) = \begin{cases} 0 & \text{if } n = 0 \\ n - 1 & \text{if } n \geq 1. \end{cases}$$

The composite $g \circ f = 1_{\mathbb{N}}$, for $g(f(n)) = g(n+1) = n$, because $n + 1 \geq 1$. On the other hand, $f \circ g \neq 1_{\mathbb{N}}$, because $f(g(0)) = f(0) = 1 \neq 0$.

Notice that f is an injection but not a surjection, and that g is a surjection but not an injection. ◂

Two strategies are now available to determine whether or not a given function is a bijection: (i) use the definitions of injection and surjection; (ii) find an inverse. For example, if $\mathbb{R}^>$ denotes the positive real numbers, let us show that the exponential function $f: \mathbb{R} \to \mathbb{R}^>$, defined by $f(x) = e^x = \sum x^n/n!$, is a bijection. It is simplest to use the (natural) logarithm $g(y) = \ln y = \int_1^y dt/t$. The usual formulas $e^{\ln y} = y$ and $\ln e^x = x$ say that both composites $f \circ g$ and $g \circ f$ are identities, and so f and g are inverse functions. Therefore, f is a bijection, for it has an inverse. (A direct proof that f is an injection would require showing that if $e^a = e^b$, then $a = b$; a direct proof showing that f is surjective would involve showing that every positive real number c has the form e^a for some a.)

Let us summarize the results just obtained.

Theorem 1.49. *If the set of all the bijections from a set X to itself is denoted by S_X, then composition of functions satisfies the following properties*:

(i) *if $f, g \in S_X$, then $f \circ g \in S_X$;*

(ii) *$h \circ (g \circ f) = (h \circ g) \circ f$ for all $f, g, h \in S_X$;*

(iii) *the identity 1_X lies in S_X, and $1_X \circ f = f = f \circ 1_X$ for every $f \in S_X$;*

(iv) *for every $f \in S_X$, there is $g \in S_X$ with $g \circ f = 1_X = f \circ g$.*

Sketch of Proof. Part (i) follows from Exercise 1.59 on page 36, which shows that the composite of two bijections is itself a bijection. The other parts of the statement have been proved above. •

If X and Y are sets, then a function $f: X \to Y$ defines a "forward motion" carrying subsets of X into subsets of Y, namely, if $S \subseteq X$, then

$$f(S) = \{y \in Y : y = f(s) \text{ for some } s \in S\},$$

and a "backward motion" carrying subsets of Y into subsets of X, namely, if $W \subseteq Y$, then

$$f^{-1}(W) = \{x \in X : f(x) \in W\}.$$

We call $f^{-1}(W)$ the ***inverse image*** of W. Formally, denote the family of all the subsets of a set X by $\mathcal{P}(X)$. If $f: X \to Y$, then there are functions

$$f_*: \mathcal{P}(X) \to \mathcal{P}(Y),$$

given by $f_*: S \mapsto f(S)$, and

$$f^*: \mathcal{P}(Y) \to \mathcal{P}(X),$$

given by $f^*: W \mapsto f^{-1}(W)$. When f is a surjection, then these motions set up a bijection between all the subsets of Y and some of the subsets of X.

Proposition 1.50. *Let X and Y be sets, and let $f : X \to Y$ be a surjection.*

(i) *If $T \subseteq S$ are subsets of X, then $f(T) \subseteq f(S)$, and if $U \subseteq V$ are subsets of Y, then $f^{-1}(U) \subseteq f^{-1}(V)$.*

(ii) *If $U \subseteq Y$, then $ff^{-1}(U) = U$.*

(iii) *The composite $f_* f^* : \mathcal{P}(Y) \to \mathcal{P}(Y) = 1_{\mathcal{P}(Y)}$, and so $f^* : W \mapsto f^{-1}(W)$ is an injection.*

(iv) *If $S \subseteq X$, then $S \subseteq f^{-1} f(S)$, but strict inclusion is possible.*

Remark. If f is not a surjection, then $W \mapsto f^{-1}(W)$ need not be an injection: There is some $y \in Y$ with $y \notin f(X)$, and $f^{-1}(\{y\}) = \varnothing = f^{-1}(\varnothing)$. ◄

Proof. (i) If $y \in f(T)$, then $y = f(t)$ for some $t \in T$. But $t \in S$, because $T \subseteq S$, and so $f(t) \in f(S)$. Therefore, $f(T) \subseteq f(S)$. The other inclusion is proved just as easily.

(ii) If $u \in U$, then f being surjective says that there is $x \in X$ with $f(x) = u$; hence, $x \in f^{-1}(U)$, and so $u = f(x) \in ff^{-1}(U)$. For the reverse inclusion, let $a \in ff^{-1}(U)$; hence, $a = f(x')$ for some $x' \in f^{-1}(U)$. But this says that $a = f(x') \in U$, as desired.

(iii) Part (ii) says that $f_* f^* = 1_{\mathcal{P}(Y)}$, and so Proposition 1.47 says that f^* is an injection.

(iv) If $s \in S$, then $f(s) \in f(S)$, and so $s \in f^{-1} f(s) \subseteq f^{-1} f(S)$.

To see that there may be strict inclusion, let $f : \mathbb{R} \to \mathbb{C}$ be given by $x \mapsto e^{2\pi i x}$. If $S = \{0\}$, then $f(S) = \{1\}$ and $f^{-1} f(\{1\}) = \mathbb{Z}$. •

In Exercise 1.68 on page 37, we will see that if $f : X \to Y$, then inverse image behaves better on subsets than does forward image; for example, $f^{-1}(S \cap T) = f^{-1}(S) \cap f^{-1}(T)$, where $S, T \subseteq Y$, but for $A, B \subseteq X$, it is possible that $f(A \cap B) \neq f(A) \cap f(B)$.

We will need cartesian products of more than two sets. One may view an element $(x_1, x_2) \in X_1 \times X_2$ as the function $f : \{1, 2\} \to X_1 \cup X_2$ with $f(i) = x_i \in X_i$ for $i = 1, 2$.

Definition. Let I be a set and let $\{X_i : i \in I\}$ be an indexed family of sets. Then the ***cartesian product*** is the set

$$\prod_{i \in I} X_i = \Big\{ f : I \to \bigcup_{i \in I} X_i : f(i) \in X_i \text{ for all } i \in I \Big\}.$$

The elements $x \in \prod_i X_i$ can be viewed as "vectors" $x = (x_i)$ whose ith coordinate is $x_i = f(i)$ for all $i \in I$. If I is finite, say, $I = \{1, 2, \ldots, n\}$, then it is not difficult to see that $\prod_i X_i = X_1 \times \cdots \times X_n$, where the latter set is defined, inductively, by

$$X_1 \times \cdots \times X_{n+1} = \big(X_1 \times \cdots \times X_n\big) \times X_{n+1}.$$

If the index set I is infinite and all the X_i are nonempty, it is not obvious that $\prod_{i \in I} X_i$ is nonempty. Indeed, this assertion is equivalent to the axiom of choice (see the Appendix).

The notion of *relation*, which generalizes that of a function, is useful.

Definition. If X and Y are sets, then a ***relation from*** X ***to*** Y is a subset $R \subseteq X \times Y$. We usually write

$$x \; R \; y$$

to denote $(x, y) \in R$. If $X = Y$, then we say that R is a ***relation on*** X.

Let us give a concrete illustration to convince the reader that this definition is reasonable. One expects that $\leq$ is a relation on $\mathbb{R}$, and let us see that it does, in fact, realize the definition of relation. Let

$$R = \{(x, y) \in \mathbb{R} \times \mathbb{R} : (x, y) \text{ lies on or above the line } y = x\}.$$

The reader should recognize that $x \; R \; y$ holds if and only if, in the usual sense, $x \leq y$.

Example 1.51.

(i) Every function $f : X \to Y$ is a relation.

(ii) Equality is a relation on any set X; it is the *diagonal*

$$\Delta_X = \{(x, x) \in X \times X\}.$$

(iii) The empty set $\varnothing$ defines a relation on any set, but it is not very interesting. ◂

Definition. A relation $x \equiv y$ on a set X is

reflexive: if $x \equiv x$ for all $x \in X$;

symmetric: if $x \equiv y$ implies $y \equiv x$ for all $x, y \in X$;

transitive: if $x \equiv y$ and $y \equiv z$ imply $x \equiv z$ for all $x, y, z \in X$.

A relation that has all three properties—reflexivity, symmetry, and transitivity—is called an ***equivalence relation***.

Example 1.52.

(i) Equality is an equivalence relation on any set X. We should regard any equivalence relation as a generalized equality.

(ii) For any integer $m \geq 0$, congruence mod m is an equivalence relation on $\mathbb{Z}$. ◂

An equivalence relation on a set X yields a family of subsets of X.

Definition. Let $\equiv$ be an equivalence relation on a set X. If $a \in X$, the ***equivalence class*** of a, denoted by $[a]$, is defined by

$$[a] = \{x \in X : x \equiv a\} \subseteq X.$$

For example, under congruence mod m, the equivalence class $[a]$ of an integer a is called its ***congruence class***.

The next lemma says that we can replace equivalence by honest equality at the cost of replacing elements by their equivalence classes.

Lemma 1.53. *If $\equiv$ is an equivalence relation on a set X, then $x \equiv y$ if and only if $[x] = [y]$.*

Proof. Assume that $x \equiv y$. If $z \in [x]$, then $z \equiv x$, and so transitivity gives $z \equiv y$; hence $[x] \subseteq [y]$. By symmetry, $y \equiv x$, and this gives the reverse inclusion $[y] \subseteq [x]$. Thus, $[x] = [y]$.

Conversely, if $[x] = [y]$, then $x \in [x]$, by reflexivity, and so $x \in [x] = [y]$. Therefore, $x \equiv y$. •

Definition. A family of subsets A_i of a set X is called ***pairwise disjoint*** if

$$A_i \cap A_j = \varnothing$$

for all $i \neq j$. A ***partition*** of a set X is a family of pairwise disjoint nonempty subsets, called ***blocks***, whose union is all of X.

Proposition 1.54. *If $\equiv$ is an equivalence relation on a set X, then the equivalence classes form a partition of X. Conversely, given a partition $\{A_i : i \in I\}$ of X, there is an equivalence relation on X whose equivalence classes are the blocks A_i.*

Proof. Assume that an equivalence relation $\equiv$ on X is given. Each $x \in X$ lies in the equivalence class $[x]$ because $\equiv$ is reflexive; it follows that the equivalence classes are nonempty subsets whose union is X. To prove pairwise disjointness, assume that $a \in [x] \cap [y]$, so that $a \equiv x$ and $a \equiv y$. By symmetry, $x \equiv a$, and so transitivity gives $x \equiv y$. Therefore, $[x] = [y]$, by the lemma, and the equivalence classes form a partition of X.

Conversely, let $\{A_i : i \in I\}$ be a partition of X. If $x, y \in X$, define $x \equiv y$ if there is $i \in I$ with both $x \in A_i$ and $y \in A_i$. It is plain that $\equiv$ is reflexive and symmetric. To see that $\equiv$ is transitive, assume that $x \equiv y$ and $y \equiv z$; that is, there are $i, j \in I$ with $x, y \in A_i$ and $y, z \in A_j$. Since $y \in A_i \cap A_j$, pairwise disjointness gives $A_i = A_j$, so that $i = j$ and $x, z \in A_i$; that is, $x \equiv z$. We have shown that $\equiv$ is an equivalence relation.

It remains to show that the equivalence classes are the A_i's. If $x \in X$, then $x \in A_i$, for some i. By definition of $\equiv$, if $y \in A_i$, then $y \equiv x$ and $y \in [x]$; hence, $A_i \subseteq [x]$. For the reverse inclusion, let $z \in [x]$, so that $z \equiv x$. There is some j with $x \in A_j$ and $z \in A_j$; thus, $x \in A_i \cap A_j$. By pairwise disjointness, $i = j$, so that $z \in A_i$, and $[x] \subseteq A_i$. Hence, $[x] = A_i$. •

Example 1.55.

(i) We have just seen that an equivalence relation can be defined on a set from a partition. Let $\mathbf{I} = [0, 1]$ be the closed unit interval, and define a partition of $\mathbf{I}$ whose blocks are the 2-point set $\{0, 1\}$ and all the 1-point sets $\{a\}$, where $0 < a < 1$. The family of all the blocks, that is, of all the equivalence classes, can be viewed as a circle, for we have identified the two endpoints of the interval.

Here is another construction of the circle, now from $\mathbb{R}$ instead of from $\mathbf{I}$. Define a relation on $\mathbb{R}$ by $a \equiv b$ if $a - b \in \mathbb{Z}$. It is easy to see that this is an equivalence relation on $\mathbb{R}$, and the equivalence class of a number a is

$$[a] = \{r \in \mathbb{R} : r = a + n \text{ for some } n \in \mathbb{Z}\}.$$

The family of all blocks is again the circle (we have identified the endpoints of any interval of length 1).

(ii) Define an equivalence relation on the square $\mathbf{I}\times\mathbf{I}$ in which the blocks are $\{(a, 0), (a, 1)\}$, one for each $a \in \mathbf{I}$, $\{(0, b), (1, b)\}$, one for each $b \in \mathbf{I}$, as well as all the singleton sets $\{(a, b)\}$ in the interior of the square. The family of all equivalence classes can be viewed as a ***torus*** (the surface of a doughnut): Identifying the left and right sides of the square gives a cylinder, and further identifying the top and bottom ends of the cylinder gives a torus. ◂

EXERCISES

1.52 Let X and Y be sets, and let $f : X \to Y$ be a function. If S is a subset of X, prove that the restriction $f|S$ is equal to the composite $f \circ i$, where $i : S \to X$ is the inclusion map.
Hint. Use Proposition 1.43.

1.53 If $f : X \to Y$ has an inverse g, show that g is a bijection.
Hint. Does g have an inverse?

1.54 Show that if $f : X \to Y$ is a bijection, then it has exactly one inverse.

1.55 Show that $f : \mathbb{R} \to \mathbb{R}$, defined by $f(x) = 3x + 5$, is a bijection, and find its inverse.

1.56 Determine whether $f : \mathbb{Q} \times \mathbb{Q} \to \mathbb{Q}$, given by

$$f(a/b, c/d) = (a + c)/(b + d),$$

is a function.

1.57 Let $X = \{x_1, \ldots, x_m\}$ and $Y = \{y_1, \ldots, y_n\}$ be finite sets. Show that there is a bijection $f : X \to Y$ if and only if $|X| = |Y|$; that is, $m = n$.
Hint. If f is a bijection, there are m distinct elements $f(x_1), \ldots, f(x_m)$ in Y, and so $m \le n$; using the bijection f^{-1} in place of f gives the reverse inequality $n \le m$.

1.58 If X and Y are finite sets with the same number of elements, show that the following conditions are equivalent for a function $f : X \to Y$:

(i) f is injective;
(ii) f is bijective;
(iii) f is surjective.

Hint. If $A \subseteq X$ and $|A| = n = |X|$, then $A = X$; after all, how many elements are in X but not in A?

1.59 Let $f : X \to Y$ and $g : Y \to Z$ be functions.
(i) If both f and g are injective, then $g \circ f$ is injective.

(ii) If both f and g are surjective, then $g \circ f$ is surjective.

(iii) If both f and g are bijective, then $g \circ f$ is bijective.

(iv) If $g \circ f$ is a bijection, prove that f is an injection and g is a surjection.

1.60 If $f\colon (-\pi/2, \pi/2) \to \mathbb{R}$ is defined by $a \mapsto \tan a$, prove that f has an inverse function g; indeed, $g = \arctan$.

1.61 If A and B are subsets of a set X, define

$$A - B = \{a \in A : a \notin B\}.$$

Prove that $A - B = A \cap B'$, where $B' = X - B$ is the ***complement*** of B; that is,

$$B' = \{x \in X : x \notin B\}.$$

1.62 Let A and B be sets, and let $a \in A$ and $b \in B$. Define their ***ordered pair*** as follows:

$$(a, b) = \{a, \{a, b\}\}.$$

If $a' \in A$ and $b' \in B$, prove that $(a', b') = (a, b)$ if and only if $a' = a$ and $b' = b$.
Hint. One of the axioms constraining the $\in$ relation is that the statement

$$a \in x \in a$$

is always false.

1.63 (i) What is wrong with the following argument, which claims to prove that a symmetric and transitive relation R on a set X is reflexive? If $x \in X$, then take $y \in X$ with $x \; R \; y$. By symmetry, we have $y \; R \; x$, and by transitivity, we have $x \; R \; x$.

(ii) Give an example of a symmetric and transitive relation on a set that is not reflexive.

1.64 (i) Let X be a set, and let $R \subseteq X \times X$. Define $\widetilde{R} = \bigcap_{R' \in \mathcal{E}} R'$, where $\mathcal{E}$ is the family of all the equivalence relations R' on X containing R. Prove that $\widetilde{R}$ is an equivalence relation on X ($\widetilde{R}$ is called the ***equivalence relation generated by*** R).

(ii) Let R be a reflexive and symmetric relation on a set X. Prove that $\widetilde{R}$, the equivalence relation generated by R, consists of all $(x, y) \in X \times X$ for which there exist finitely many $(x, y) \in R$, say, $(x_1, y_1), \ldots, (x_n, y_n)$, with $x = x_1$, $y_n = y$, and $y_i = x_{i+1}$ for all $i \geq 1$.

1.65 Let $X = \{(a, b) : a, b \in \mathbb{Z}$ and $b \neq 0\}$. Prove that the relation on X, defined by $(a, b) \equiv (c, d)$ if $ad = bc$, is an equivalence relation on X. What is the equivalence class of $(1, 2)$?

1.66 Define a relation on $\mathbb{C}$ by $z \equiv w$ if $|z| = |w|$. Prove that this is an equivalence relation on $\mathbb{C}$ whose equivalence classes are the origin and the circles with center the origin.

1.67 (i) Let $f\colon X \to Y$ be a function (where X and Y are sets). Prove that the relation on X, defined by $x \equiv x'$ if $f(x) = f(x')$, is an equivalence relation.

(ii) Define $f\colon \mathbb{R} \to S^1$, where $S^1 \subseteq \mathbb{C}$ is the unit circle, by $f(x) = e^{2\pi i x}$. What is the equivalence class of 0 under the equivalence relation in part (i)?

1.68 Let $f\colon X \to Y$ be a function and let $V, W \subseteq Y$.

(i) Prove that

$$f^{-1}(V \cap W) = f^{-1}(V) \cap f^{-1}(W) \quad \text{and} \quad f^{-1}(V \cup W) = f^{-1}(V) \cup f^{-1}(W).$$

(ii) Prove that $f(V \cup W) = f(V) \cup f(W)$.

(iii) Give an example showing that $f(V \cap W) \neq f(V) \cap f(W)$.

(iv) Prove that $f^{-1}(W') = (f^{-1}(W))'$, where $W' = \{y \in Y : y \notin W\}$ is the complement of W, and give an example of a function f such that $f(S') \neq (f(S))'$ for some $S \subseteq X$.

2

Groups I

2.1 Introduction

One of the major open problems, following the discovery of the cubic and quartic formulas in the 1500s, was to find a formula for the roots of polynomials of higher degree, and it remained open for almost 300 years. For about the first 100 years, mathematicians reconsidered what *number* means, for understanding the cubic formula forced such questions as whether negative numbers are numbers and whether complex numbers are legitimate entities as well. By 1800, P. Ruffini claimed that there is no quintic formula (which has the same form as the quadratic, cubic, and quartic formulas; that is, it uses only arithmetic operations and nth roots), but his contemporaries did not accept his proof (his ideas were, in fact, correct, but his proof had gaps). In 1815, A. L. Cauchy introduced the multiplication of permutations and proved basic properties of what we call the symmetric group S_n; for example, he introduced the cycle notation and proved the unique factorization of permutations into disjoint cycles. In 1824, N. Abel (1802-1829) gave an acceptable proof that there is no quintic formula; in his proof, Abel constructed permutations of the roots of a quintic, using certain rational functions introduced by J. L. Lagrange in 1770. E. Galois (1811–1832), the young wizard who was killed before his 21st birthday, modified the rational functions but, more important, he saw that the key to understanding the problem involved what he called *groups*: subsets of S_n that are closed under multiplication – in our language, *subgroups* of S_n. To each polynomial $f(x)$, he associated such a group, nowadays called the *Galois group* of $f(x)$. He recognized conjugation, normal subgroups, quotient groups, and simple groups, and he proved, in our language, that a polynomial (over a field of characteristic 0) has a formula for its roots, analogous to the quadratic formula, if and only if its Galois group is a *solvable group* (solvability being a property generalizing commutativity). A good case can be made that Galois was one of the most important founders of modern algebra. For an excellent account of the history of this problem we recommend the book, *Galois' Theory of Algebraic Equations*, by J.-P. Tignol.

Along with results usually not presented in a first course, this chapter will also review some familiar results whose proofs will only be sketched.

2.2 PERMUTATIONS

For Galois, groups consisted of certain permutations (of the roots of a polynomial), and groups of permutations remain important today.

Definition. A ***permutation*** of a set X is a bijection from X to itself.

In high school mathematics, a permutation of a set X is defined as a rearrangement of its elements. For example, there are six rearrangements of $X = \{1, 2, 3\}$:

$$123; \quad 132; \quad 213; \quad 231; \quad 312; \quad 321.$$

Now let $X = \{1, 2, \ldots, n\}$. A *rearrangement* is a list, with no repetitions, of all the elements of X. All we can do with such lists is count them, and there are exactly $n!$ permutations of the n-element set X.

Now a rearrangement $i_1, i_2, \ldots, i_n$ of X determines a function $\alpha \colon X \to X$, namely, $\alpha(1) = i_1, \alpha(2) = i_2, \ldots, \alpha(n) = i_n$. For example, the rearrangement 213 determines the function α with $\alpha(1) = 2, \alpha(2) = 1$, and $\alpha(3) = 3$. We use a two-rowed notation to denote the function corresponding to a rearrangement; if $\alpha(j)$ is the jth item on the list, then

$$\alpha = \begin{pmatrix} 1 & 2 & \ldots & j & \ldots & n \\ \alpha(1) & \alpha(2) & \ldots & \alpha(j) & \ldots & \alpha(n) \end{pmatrix}.$$

That a list contains *all* the elements of X says that the corresponding function α is surjective, for the bottom row is $\operatorname{im}\alpha$; that there are no repetitions on the list says that distinct points have distinct values; that is, α is injective. Thus, each list determines a bijection $\alpha \colon X \to X$; that is, each rearrangement determines a permutation. Conversely, every permutation α determines a rearrangement, namely, the list $\alpha(1), \alpha(2), \ldots, \alpha(n)$ displayed as the bottom row. Therefore, rearrangement and permutation are simply different ways of describing the same thing. The advantage of viewing permutations as functions, however, is that they can now be composed and, by Exercise 1.59 on page 36, their composite is also a permutation.

Definition. The family of all the permutations of a set X, denoted by S_X, is called the ***symmetric group*** on X. When $X = \{1, 2, \ldots, n\}$, S_X is usually denoted by S_n, and it is called the ***symmetric group on*** n ***letters***.

Let us simplify notation by writing $\beta\alpha$ instead of $\beta \circ \alpha$ and (1) instead of 1_X.

Notice that composition in S_3 is not commutative. Aside from being cumbersome, there is a major problem with the two-rowed notation for permutations. It hides the answers to elementary questions such as, Do two permutations commute? Is the square of a permutation the identity? The special permutations introduced next will remedy this defect.

Definition. Let $i_1, i_2, \ldots, i_r$ be distinct integers in $\{1, 2, \ldots, n\}$. If $\alpha \in S_n$ fixes the other integers (if any) and if

$$\alpha(i_1) = i_2, \alpha(i_2) = i_3, \ \ldots, \alpha(i_{r-1}) = i_r, \alpha(i_r) = i_1,$$

then α is called an ***r-cycle***. We also say that α is a cycle of ***length*** r, and we denote it by

$$\alpha = (i_1 \ i_2 \ \ldots \ i_r).$$

A 2-cycle interchanges i_1 and i_2 and fixes everything else; 2-cycles are also called ***transpositions***. A 1-cycle is the identity, for it fixes every i; thus, all 1-cycles are equal: $(i) = (1)$ for all i.

The term *cycle* comes from the Greek word for circle. Picture the cycle $(i_1 \ i_2 \ \ldots \ i_r)$ as a clockwise rotation of the circle, as in Figure 2.1.

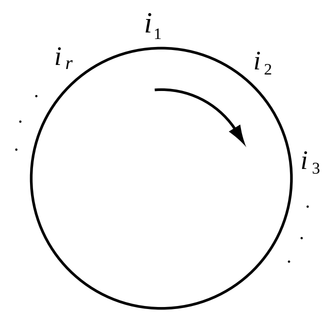


Figure 2.1

Any i_j can be taken as the "starting point," and so there are r different cycle notations for any r-cycle:

$$(i_1 \ i_2 \ \ldots \ i_r) = (i_2 \ i_3 \ \ldots \ i_r \ i_1) = \cdots = (i_r \ i_1 \ i_2 \ \ldots \ i_{r-1}).$$

Let us now give an ***algorithm*** to factor a permutation into a product of cycles. For example, take

$$\alpha = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\ 6 & 4 & 7 & 2 & 5 & 1 & 8 & 9 & 3 \end{pmatrix}.$$

Begin by writing "(1." Now $\alpha \colon 1 \mapsto 6$, so write "(1 6." Next, $\alpha \colon 6 \mapsto 1$, and so the parentheses close: α begins "(1 6)." The first number not having appeared is 2, and so we write "(1 6)(2." Now $\alpha \colon 2 \mapsto 4$, so we write "(1 6)(2 4." Since $\alpha \colon 4 \mapsto 2$, the parentheses close once again, and we write "(1 6)(2 4)." The smallest remaining number is 3; now $3 \mapsto 7$, $7 \mapsto 8$, $8 \mapsto 9$, and $9 \mapsto 3$; this gives the 4-cycle (3 7 8 9). Finally, $\alpha(5) = 5$; we claim that

$$\alpha = (1\ 6)(2\ 4)(3\ 7\ 8\ 9)(5).$$

Since multiplication in S_n is composition of functions, our claim is that

$$\alpha(i) = [(1\ 6)(2\ 4)(3\ 7\ 8\ 9)(5)](i)$$

for every i between 1 and 9 [after all, two functions f and g are equal if and only if $f(i) = g(i)$ for every i in their domain]. The right side is the composite $\beta\gamma\delta$, where $\beta = (1\ 6)$, $\gamma = (2\ 4)$, and $\delta = (3\ 7\ 8\ 9)$ [we may ignore the 1-cycle (5) when we are evaluating, for it is the identity function]. Now $\alpha(1) = 6$; let us evaluate the composite on the right when $i = 1$.

$$\begin{aligned}
\beta\gamma\delta(1) &= \beta(\gamma(\delta(1))) \\
&= \beta(\gamma(1)) && \delta = (3\ 7\ 8\ 9) \text{ fixes } 1 \\
&= \beta(1) && \gamma = (2\ 4) \text{ fixes } 1 \\
&= 6 && \beta = (1\ 6).
\end{aligned}$$

Similarly, $\alpha(i) = \beta\gamma\delta(i)$ for every i, proving the claim.

We multiply permutations from right to left, because multiplication here is composite of functions; that is, to evaluate $\alpha\beta(1)$, we compute $\alpha(\beta(1))$. Here is another example: Let us compute the product

$$\sigma = (1\ 2)(1\ 3\ 4\ 2\ 5)(2\ 5\ 1\ 3)$$

in S_5. To find the two-rowed notation for σ, evaluate, starting with the cycle on the right:

$$\begin{aligned}
\sigma &: 1 \mapsto 3 \mapsto 4 \mapsto 4; \\
\sigma &: 2 \mapsto 5 \mapsto 1 \mapsto 2; \\
\sigma &: 3 \mapsto 2 \mapsto 5 \mapsto 5; \\
\sigma &: 4 \mapsto 4 \mapsto 2 \mapsto 1; \\
\sigma &: 5 \mapsto 1 \mapsto 3 \mapsto 3.
\end{aligned}$$

Thus,[1]

$$\sigma = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 4 & 2 & 5 & 1 & 3 \end{pmatrix}.$$

The algorithm given earlier, when applied to this two-rowed notation for σ, now gives

$$\sigma = (1\ 4)(2)(5\ 3).$$

In the factorization of a permutation into cycles, given by the preceding algorithm, we note that the family of cycles is *disjoint* in the following sense.

[1]There are authors who multiply permutations differently, so that their $\alpha \circ \beta$ is our $\beta \circ \alpha$. This is a consequence of their putting "functions on the right": Instead of writing $\alpha(i)$ as we do, they write $(i)\alpha$. Consider the composite of permutations α and β in which we first apply β and then apply α. We write $i \mapsto \beta(i) \mapsto \alpha(\beta(i))$. In the right-sided notation, $i \mapsto (i)\beta \mapsto ((i)\beta)\alpha$. Thus, the notational switch causes a switch in the order of multiplication.

Definition. Two permutations $\alpha, \beta \in S_n$ are ***disjoint*** if every i moved by one is fixed by the other: If $\alpha(i) \neq i$, then $\beta(i) = i$, and if $\beta(j) \neq j$, then $\alpha(j) = j$. A family $\beta_1 \ldots, \beta_t$ of permutations is ***disjoint*** if each pair of them is disjoint.

Lemma 2.1. *Disjoint permutations $\alpha, \beta \in S_n$ commute.*

Proof. It suffices to prove that if $1 \leq i \leq n$, then $\alpha\beta(i) = \beta\alpha(i)$. If β moves i, say, $\beta(i) = j \neq i$, then β also moves j [otherwise, $\beta(j) = j$ and $\beta(i) = j$ contradicts β's being an injection]; since α and β are disjoint, $\alpha(i) = i$ and $\alpha(j) = j$. Hence $\beta\alpha(i) = j = \alpha\beta(i)$. The same conclusion holds if α moves i. Finally, it is clear that $\alpha\beta(i) = \beta\alpha(i)$ if both α and β fix i. •

Proposition 2.2. *Every permutation $\alpha \in S_n$ is either a cycle or a product of disjoint cycles.*

Proof. The proof is by induction on the number k of points moved by α. The base step $k = 0$ is true, for now α is the identity, which is a 1-cycle.

If $k > 0$, let i_1 be a point moved by α. Define $i_2 = \alpha(i_1)$, $i_3 = \alpha(i_2), \ldots, i_{r+1} = \alpha(i_r)$, where r is the smallest integer for which $i_{r+1} \in \{i_1, i_2, \ldots, i_r\}$ (since there are only n possible values, the list $i_1, i_2, i_3, \ldots, i_k, \ldots$ must eventually have a repetition). We claim that $\alpha(i_r) = i_1$. Otherwise, $\alpha(i_r) = i_j$ for some $j \geq 2$; but $\alpha(i_{j-1}) = i_j$, and this contradicts the hypothesis that α is an injection. Let σ be the r-cycle $(i_1\ i_2\ i_3\ \ldots\ i_r)$. If $r = n$, then $\alpha = \sigma$. If $r < n$, then σ fixes each point in Y, where Y consists of the remaining $n-r$ points, while $\alpha(Y) = Y$. Define α' to be the permutation with $\alpha'(i) = \alpha(i)$ for $i \in Y$ that fixes all $i \notin Y$, and note that

$$\alpha = \sigma\alpha'.$$

The inductive hypothesis gives $\alpha' = \beta_1 \cdots \beta_t$, where $\beta_1, \ldots, \beta_t$ are disjoint cycles. Since σ and α' are disjoint, $\alpha = \sigma\beta_1 \cdots \beta_t$ is a product of disjoint cycles. •

Usually we suppress the 1-cycles in this factorization [for 1-cycles equal the identity (1)]. However, a factorization of α in which we display one 1-cycle for each i fixed by α, if any, will arise several times.

Definition. A ***complete factorization*** of a permutation α is a factorization of α into disjoint cycles that contains exactly one 1-cycle (i) for every i fixed by α.

For example, the complete factorization of the 3-cycle $\alpha = (1\ 3\ 5)$ in S_5 is $\alpha = (1\ 3\ 5)(2)(4)$.

There is a relation between an r-cycle $\beta = (i_1\ i_2\ \ldots\ i_r)$ and its ***powers*** β^k, where β^k denotes the composite of β with itself k times. Note that $i_2 = \beta(i_1)$, $i_3 = \beta(i_2) = \beta(\beta(i_1)) = \beta^2(i_1)$, $i_4 = \beta(i_3) = \beta(\beta^2(i_1)) = \beta^3(i_1)$, and, more generally,

$$i_{k+1} = \beta^k(i_1)$$

for all $k < r$.

Theorem 2.3. *Let $\alpha \in S_n$ and let $\alpha = \beta_1 \cdots \beta_t$ be a complete factorization into disjoint cycles. This factorization is unique except for the order in which the cycles occur.*

Sketch of Proof. Since every complete factorization of α has exactly one 1-cycle for each i fixed by α, it suffices to consider (not complete) factorizations into disjoint cycles of length ≥ 2. Let $\alpha = \gamma_1 \cdots \gamma_s$ be a second such factorization of α into disjoint cycles.

The theorem is proved by induction on ℓ, the larger of t and s. The inductive step begins by noting that if β_t moves i_1, then $\beta_t^k(i_1) = \alpha^k(i_1)$ for all $k \geq 1$. Some γ_j must also move i_1 and, since disjoint cycles commute, we may assume that γ_s moves i_1. It follows that $\beta_t = \gamma_s$; right multiplying by β_t^{-1} gives $\beta_1 \cdots \beta_{t-1} = \gamma_1 \cdots \gamma_{s-1}$. •

Every permutation is a bijection; how do we find its inverse? In the pictorial representation of a cycle β as a clockwise rotation of a circle, the inverse β^{-1} is just a counterclockwise rotation. The proof of the next proposition is straightforward.

Proposition 2.4.

(i) *The inverse of the cycle $\alpha = (i_1\ i_2\ \ldots\ i_r)$ is the cycle $(i_r\ i_{r-1} \ldots\ i_1)$:*

$$(i_1\ i_2\ \ldots\ i_r)^{-1} = (i_r\ i_{r-1} \ldots\ i_1).$$

(ii) *If $\gamma \in S_n$ and $\gamma = \beta_1 \cdots \beta_k$, then*

$$\gamma^{-1} = \beta_k^{-1} \cdots \beta_1^{-1}.$$

Definition. Two permutations $\alpha, \beta \in S_n$ have the ***same cycle structure*** if their complete factorizations have the same number of r-cycles for each r.

According to Exercise 2.4 on page 50, there are

$$(1/r)[n(n-1)\cdots(n-r+1)]$$

r-cycles in S_n. This formula can be used to count the number of permutations having any given cycle structure if we are careful about factorizations having several cycles of the same length. For example, the number of permutations in S_4 of the form $(a\ b)(c\ d)$ is $\frac{1}{2}\big[\frac{1}{2}(4 \times 3)\big] \times \big[\frac{1}{2}(2 \times 1)\big] = 3$, the "extra" factor $\frac{1}{2}$ occurring so that we do not count $(a\ b)(c\ d) = (c\ d)(a\ b)$ twice.

Example 2.5.

(i) The types of permutations in $G = S_4$ are counted in Table 2.1.

Cycle Structure	Number
(1)	1
(1 2)	6
(1 2 3)	8
(1 2 3 4)	6
(1 2)(3 4)	3
	24

Table 2.1. Permutations in S_4

(ii) The types of permutations in $G = S_5$ are counted in Table 2.2.

Cycle Structure	Number
(1)	1
(1 2)	10
(1 2 3)	20
(1 2 3 4)	30
(1 2 3 4 5)	24
(1 2)(3 4 5)	20
(1 2)(3 4)	15
	120

◂

Table 2.2. Permutations in S_5

Here is a computational aid. We illustrate its statement in the following example before stating the general result.

Example 2.6.
If $\gamma = (1\ 3)(2\ 4\ 7)(5)(6)$ and $\alpha = (2\ 5\ 6)(1\ 4\ 3)$, then

$$\alpha\gamma\alpha^{-1} = (4\ 1)(5\ 3\ 7)(6)(2) = (\alpha 1\ \alpha 3)(\alpha 2\ \alpha 4\ \alpha 7)(\alpha 5)(\alpha 6). \quad ◂$$

Lemma 2.7. *If $\gamma, \alpha \in S_n$, then $\alpha\gamma\alpha^{-1}$ has the same cycle structure as γ. In more detail, if the complete factorization of γ is*

$$\gamma = \beta_1\beta_2 \cdots (i_1\ i_2\ \ldots) \cdots \beta_t,$$

then $\alpha\gamma\alpha^{-1}$ is the permutation that is obtained from γ by applying α to the symbols in the cycles of γ.

Proof. The idea of the proof is that $\gamma\alpha\gamma^{-1}\colon \gamma(i_1) \mapsto i_1 \mapsto i_2 \mapsto \gamma(i_2)$. Let σ denote the permutation defined in the statement.

If γ fixes i, then σ fixes $\alpha(i)$, for the definition of σ says that $\alpha(i)$ lives in a 1-cycle in the factorization of σ. On the other hand, $\alpha\gamma\alpha^{-1}$ also fixes $\alpha(i)$:

$$\alpha\gamma\alpha^{-1}(\alpha(i)) = \alpha\gamma(i) = \alpha(i),$$

because γ fixes i.

Assume that γ moves a symbol i_1, say, $\gamma(i_1) = i_2$, so that one of the cycles in the complete factorization of γ is

$$(i_1\ i_2\ \ldots).$$

By the definition of σ, one of its cycles is

$$(k\ \ell\ \ldots),$$

where $\alpha(i_1) = k$ and $\alpha(i_2) = \ell$; hence, $\sigma\colon k \mapsto \ell$. But $\alpha\gamma\alpha^{-1}\colon k \mapsto i_1 \mapsto i_2 \mapsto \ell$, and so $\alpha\gamma\alpha^{-1}(k) = \sigma(k)$. Therefore, σ and $\alpha\gamma\alpha^{-1}$ agree on all symbols of the form $k = \alpha(i_1)$. Since α is surjective, every k is of this form, and so $\sigma = \alpha\gamma\alpha^{-1}$. •

Example 2.8.
In this example, we illustrate that the converse of Lemma 2.7 is true; the next theorem will prove it in general. In S_5, place the complete factorization of a 3-cycle β over that of a 3-cycle γ, and define α to be the downward function. For example, if

$$\beta = (1\ 2\ 3)(4)(5)$$
$$\gamma = (5\ 2\ 4)(1)(3),$$

then

$$\alpha = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 5 & 2 & 4 & 1 & 3 \end{pmatrix},$$

and so $\alpha = (1\ 5\ 3\ 4)$. Now $\alpha \in S_5$ and

$$\gamma = (\alpha 1\ \alpha 2\ \alpha 3),$$

so that $\gamma = \alpha\beta\alpha^{-1}$, by Lemma 2.7. Note that rewriting the cycles of β, for example, as $\beta = (1\ 2\ 3)(5)(4)$, gives another choice for α. ◀

Theorem 2.9. *Permutations γ and σ in S_n have the same cycle structure if and only if there exists $\alpha \in S_n$ with $\sigma = \alpha\gamma\alpha^{-1}$.*

Sketch of Proof. Sufficiency was just proved in Lemma 2.7. For the converse, place one complete factorization over the other so that each cycle below is under a cycle above of the same length:

$$\gamma = \delta_1\delta_2 \cdots (i_1\ i_2 \cdots) \cdots \delta_t$$
$$\alpha\gamma\alpha^{-1} = \eta_1\eta_2 \cdots (k\ \ell \cdots) \cdots \eta_t.$$

Now define α to be the "downward" function, as in the example; hence, $\alpha(i_1) = k$, $\alpha(i_2) = \ell$, and so forth. Note that α is a permutation, for there are no repetitions of symbols in the factorization of γ (the cycles η are disjoint). It now follows from the lemma that $\sigma = \alpha\gamma\alpha^{-1}$. •

There is another useful factorization of a permutation.

Proposition 2.10. *If $n \geq 2$, then every $\alpha \in S_n$ is a product of transpositions.*

Sketch of Proof. In light of Proposition 2.2, it suffices to factor an r-cycle β into a product of transpositions, and this is done as follows:

$$\beta = (1\ 2\ \ldots\ r) = (1\ r)(1\ r-1) \cdots (1\ 3)(1\ 2). \quad \bullet$$

Every permutation can thus be realized as a sequence of interchanges, but such a factorization is not as nice as the factorization into disjoint cycles. First, the transpositions occurring need not commute: $(1\ 2\ 3) = (1\ 3)(1\ 2) \neq (1\ 2)(1\ 3)$; second, neither the factors themselves nor the number of factors are uniquely determined. For example, here are some factorizations of (1 2 3) in S_4:

$$\begin{aligned}(1\ 2\ 3) &= (1\ 3)(1\ 2)\\ &= (2\ 3)(1\ 3)\\ &= (1\ 3)(4\ 2)(1\ 2)(1\ 4)\\ &= (1\ 3)(4\ 2)(1\ 2)(1\ 4)(2\ 3)(2\ 3).\end{aligned}$$

Is there any uniqueness at all in such a factorization? We now prove that the parity of the number of factors is the same for all factorizations of a permutation α; that is, the number of transpositions is always even or always odd (as suggested by the factorizations of $\alpha = (1\ 2\ 3)$ displayed above).

Example 2.11.
The ***15-puzzle*** has a ***starting position*** that is a 4×4 array of the numbers between 1 and 15 and a symbol #, which we interpret as "blank." For example, consider the following starting position:

3	15	4	8
10	11	1	9
2	5	13	12
6	7	14	#

A ***simple move*** interchanges the blank with a symbol adjacent to it; for example, there are two beginning simple moves for this starting position: Either interchange # and 14 or interchange # and 12. We win the game if, after a sequence of simple moves, the starting position is transformed into the standard array 1, 2, 3, . . ., 15, #.

To analyze this game, note that the given array is really a permutation $\alpha \in S_{16}$ (if we now call the blank 16 instead of #). More precisely, if the spaces are labeled 1 through 16, then $\alpha(i)$ is the symbol occupying the ith square. For example, the given starting position is

$$\begin{pmatrix}1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 & 13 & 14 & 15 & 16\\ 3 & 15 & 4 & 8 & 10 & 11 & 1 & 9 & 2 & 5 & 13 & 12 & 6 & 7 & 14 & 16\end{pmatrix}.$$

Each simple move is a special kind of transposition, namely, one that moves 16 (remember that the blank is now 16). Moreover, performing a simple move (corresponding to a special transposition τ) from a given position (corresponding to a permutation β) yields a new position corresponding to the permutation $\tau\beta$. For example, if α is the position above and τ is the transposition interchanging 14 and 16, then $\tau\alpha(16) = \tau(16) = 14$ and $\tau\alpha(15) = \tau(14) = 16$, while $\tau\alpha(i) = i$ for all other i. That is, the new configuration has all the numbers in their original positions except for 14 and 16 being interchanged. To win the

game, we need special transpositions $\tau_1, \tau_2, \ldots, \tau_m$ so that

$$\tau_m \cdots \tau_2 \tau_1 \alpha = (1).$$

It turns out that there are some choices of α for which the game can be won, but there are others for which it cannot be won, as we shall see in Example 2.15. ◂

Definition. A permutation $\alpha \in S_n$ is ***even*** if it can be factored into a product of an even number of transpositions; otherwise, α is ***odd***. The ***parity*** of a permutation is whether it is even or odd.

It is easy to see that $(1\ 2\ 3)$ and (1) are even permutations, for there are factorization $(1\ 2\ 3) = (1\ 3)(1\ 2)$ and $(1) = (1\ 2)(1\ 2)$ having two transpositions. On the other hand, we do not yet have any examples of odd permutations! If α is a product of an odd number of transpositions, perhaps it also has some other factorization into an even number of transpositions. The definition of odd permutation α, after all, says that there is no factorization of α into an even number of transpositions.

Definition. If $\alpha \in S_n$ and $\alpha = \beta_1 \cdots \beta_t$ is a complete factorization into disjoint cycles, then ***signum*** α is defined by

$$\operatorname{sgn}(\alpha) = (-1)^{n-t}.$$

Theorem 2.3 shows that sgn is a (well-defined) function, for the number t is uniquely determined by α. Notice that $\operatorname{sgn}(\varepsilon) = 1$ for every 1-cycle ε because $t = n$. If τ is a transposition, then it moves two numbers, and it fixes each of the $n-2$ other numbers; therefore, $t = (n-2)+1 = n-1$, and so $\operatorname{sgn}(\tau) = (-1)^{n-(n-1)} = -1$.

Theorem 2.12. *For all $\alpha, \beta \in S_n$,*

$$\operatorname{sgn}(\alpha\beta) = \operatorname{sgn}(\alpha)\operatorname{sgn}(\beta).$$

Sketch of Proof. If $k, \ell \geq 0$ and the letters a, b, c_i, d_j are all distinct, then

$$(a\ b)(a\ c_1\ \ldots\ c_k\ b\ d_1\ \ldots\ d_\ell) = (a\ c_1\ \ldots\ c_k)(b\ d_1\ \ldots\ d_\ell);$$

multiplying this equation on the left by $(a\ b)$ gives

$$(a\ b)(a\ c_1\ \ldots\ c_k)(b\ d_1\ \ldots\ d_\ell) = (a\ c_1\ \ldots\ c_k\ b\ d_1\ \ldots\ d_\ell).$$

These equations are used to prove that $\operatorname{sgn}(\tau\alpha) = -\operatorname{sgn}(\alpha)$ for every $\alpha \in S_n$, where τ is the transposition $(a\ b)$. If $\alpha \in S_n$ has a factorization $\alpha = \tau_1 \cdots \tau_m$, where each τ_i is a transposition, we now prove, by induction on m, that $\operatorname{sgn}(\alpha\beta) = \operatorname{sgn}(\alpha)\operatorname{sgn}(\beta)$ for every $\beta \in S_n$. •

Theorem 2.13.

(i) *Let* $\alpha \in S_n$; *if* $\operatorname{sgn}(\alpha) = 1$, *then* α *is even, and if* $\operatorname{sgn}(\alpha) = -1$, *then* α *is odd.*

(ii) *A permutation* α *is odd if and only if it is a product of an odd number of transpositions.*

Proof. (i) If $\alpha = \tau_1 \cdots \tau_q$ is a factorization of α into transpositions, then Theorem 2.12 gives $\operatorname{sgn}(\alpha) = \operatorname{sgn}(\tau_1) \cdots \operatorname{sgn}(\tau_q) = (-1)^q$. Thus, if $\operatorname{sgn}(\alpha) = 1$, then q must always be even, and if $\operatorname{sgn}(\alpha) = -1$, then q must always be odd.

(ii) If α is odd, then α is not even, and so $\operatorname{sgn}(\alpha) \neq 1$; that is, $\operatorname{sgn}(\alpha) = -1$. Now $\alpha = \tau_1 \cdots \tau_q$, where the τ_i are transpositions, so that $\operatorname{sgn}(\alpha) = -1 = (-1)^q$; hence, q is odd (we have proved more; every factorization of α into transpositions has an odd number of factors). Conversely, if $\alpha = \tau_1 \cdots \tau_q$ is a product of transpositions with q odd, then $\operatorname{sgn}(\alpha) = -1$; therefore, α is not even and, hence, α is odd. •

Corollary 2.14. *Let* $\alpha, \beta \in S_n$. *If* α *and* β *have the same parity, then* $\alpha\beta$ *is even, while if* α *and* β *have distinct parity, then* $\alpha\beta$ *is odd.*

Example 2.15.

An analysis of the 15-puzzle in Example 2.11 shows that if $\alpha \in S_{16}$ is the starting position, then the game can be won if and only if α is an even permutation that fixes 16. For a proof of this, we refer the reader to McCoy–Janusz, *Introduction to Modern Algebra*, pages 229-234. The proof in one direction is fairly clear, however. The blank 16 starts in position 16. Each simple move takes 16 up, down, left, or right. Thus, the total number m of moves is $u + d + l + r$, where u is the number of up moves, and so on. If 16 is to return home, each one of these must be undone: There must be the same number of up moves as down moves (i.e., $u = d$) and the same number of left moves as right moves (i.e., $r = l$). Thus, the total number of moves is even: $m = 2u + 2r$. That is, if $\tau_m \cdots \tau_1 \alpha = (1)$, then m is even; hence, $\alpha = \tau_1 \cdots \tau_m$ (because $\tau^{-1} = \tau$ for every transposition τ), and so α is an even permutation. Armed with this theorem, we see that if the starting position α is odd, the game starting with α cannot be won. In Example 2.11,

$$\alpha = (1\ 3\ 4\ 8\ 9\ 2\ 15\ 14\ 7)(5\ 10)(6\ 11\ 13)(12)(16)$$

[(12) and (16) are 1-cycles]. Now $\operatorname{sgn}(\alpha) = (-1)^{16-5} = -1$, so that α is an odd permutation. Therefore, it is impossible to win this game. ◂

EXERCISES

2.1 Find $\operatorname{sgn}(\alpha)$ and α^{-1}, where

$$\alpha = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\ 9 & 8 & 7 & 6 & 5 & 4 & 3 & 2 & 1 \end{pmatrix}.$$

2.2 If $\alpha \in S_n$, prove that $\operatorname{sgn}(\alpha^{-1}) = \operatorname{sgn}(\alpha)$.

2.3 If $\sigma \in S_n$ fixes some j, where $1 \le j \le n$ [that is, $\sigma(j) = j$], define $\sigma' \in S_{n-1}$ by $\sigma'(i) = \sigma(i)$ for all $i \ne j$. Prove that

$$\operatorname{sgn}(\sigma') = \operatorname{sgn}(\sigma).$$

Hint. Use the complete factorizations of σ and of σ'.

2.4 If $1 \le r \le n$, show that there are

$$\tfrac{1}{r}[n(n-1)\cdots(n-r+1)]$$

r-cycles in S_n.
Hint. There are r cycle notations for any r-cycle.

2.5 (i) If α is an r-cycle, show that $\alpha^r = (1)$.
Hint. If $\alpha = (i_0 \dots i_{r-1})$, show that $\alpha^k(i_0) = i_k$.
(ii) If α is an r-cycle, show that r is the smallest positive integer k such that $\alpha^k = (1)$.
Hint. Use Proposition 2.2.

2.6 Show that an r-cycle is an even permutation if and only if r is odd.

2.7 Given $X = \{1, 2, \dots, n\}$, let us call a permutation τ of X an ***adjacency*** if it is a transposition of the form $(i\ i+1)$ for $i < n$.
(i) Prove that every permutation in S_n, for $n \ge 2$, is a product of adjacencies.
(ii) If $i < j$, prove that $(i\ j)$ is a product of an odd number of adjacencies.
Hint. Use induction on $j - i$.

2.8 Define $f\colon \{0, 1, 2, \dots, 10\} \to \{0, 1, 2, \dots, 10\}$ by

$$f(n) = \text{the remainder after dividing } 4n^2 - 3n^7 \text{ by } 11.$$

(i) Show that f is a permutation.[2]
(ii) Compute the parity of f.
(iii) Compute the inverse of f.

2.9 If α is an r-cycle and $1 < k < r$, is α^k an r-cycle?

2.10 (i) Prove that if α and β are (not necessarily disjoint) permutations that commute, then $(\alpha\beta)^k = \alpha^k\beta^k$ for all $k \ge 1$.
Hint. First show that $\beta\alpha^k = \alpha^k\beta$ by induction on k.
(ii) Give an example of two permutations α and β for which $(\alpha\beta)^2 \ne \alpha^2\beta^2$.

2.11 (i) Prove, for all i, that $\alpha \in S_n$ moves i if and only if α^{-1} moves i.
(ii) Prove that if $\alpha, \beta \in S_n$ are disjoint and if $\alpha\beta = (1)$, then $\alpha = (1)$ and $\beta = (1)$.

2.12 Prove that the number of even permutations in S_n is $\tfrac{1}{2}n!$.
Hint. Let $\tau = (1\ 2)$, and define $f\colon A_n \to O_n$, where A_n is the set of all even permutations in S_n and O_n is the set of all odd permutations, by

$$f\colon \alpha \mapsto \tau\alpha.$$

Show that f is a bijection, so that $|A_n| = |O_n|$ and, hence, $|A_n| = \tfrac{1}{2}n!$.

---

[2] If k is a finite field, then a polynomial $f(x)$ with coefficients in k is called a ***permutation polynomial*** if the evaluation function $f\colon k \to k$, defined by $a \mapsto f(a)$, is a permutation of k. A theorem of Hermite and Dickson characterizes permutation polynomials (see Lidl–Niederreiter, *Introduction to Finite Fields and Their Applications*).

2.13 (i) How many permutations in S_5 commute with $\alpha = (1\ 2\ 3)$, and how many *even* permutations in S_5 commute with α?

Hint. There are 6 permutations in S_5 commuting with α, only 3 of which are even.

(ii) Same questions for $(1\ 2)(3\ 4)$.

Hint. There are 8 permutations in S_4 commuting with $(1\ 2)(3\ 4)$, and only 4 of them are even.

2.14 Give an example of $\alpha, \beta, \gamma \in S_5$, with $\alpha \neq (1)$, such that $\alpha\beta = \beta\alpha$, $\alpha\gamma = \gamma\alpha$ and $\beta\gamma \neq \gamma\beta$.

2.15 If $n \geq 3$, show that if $\alpha \in S_n$ commutes with every $\beta \in S_n$, then $\alpha = (1)$.

2.16 If $\alpha = \beta_1 \cdots \beta_m$ is a product of disjoint cycles, prove that $\gamma = \beta_1^{e_1} \cdots \beta_m^{e_m} \delta$ commutes with α, where $e_i \geq 0$ for all i, and δ is disjoint from α.

2.3 GROUPS

Since Galois's time, groups have arisen in many areas of mathematics other than the study of roots of polynomials, for they are the way to describe the notion of symmetry, as we shall see.

The essence of a "product" is that two things are combined to form a third thing of the same kind. For example, ordinary multiplication, addition, and subtraction combine two numbers to give another number, while composition combines two permutations to give another permutation.

Definition. A ***binary operation*** on a set G is a function

$$* : G \times G \to G.$$

In more detail, a binary operation assigns an element $*(x, y)$ in G to each ordered pair (x, y) of elements in G. It is more natural to write $x * y$ instead of $*(x, y)$; thus, composition of functions is the function $(g, f) \mapsto g \circ f$; multiplication, addition, and subtraction are, respectively, the functions $(x, y) \mapsto xy$, $(x, y) \mapsto x + y$, and $(x, y) \mapsto x - y$. The examples of composition and subtraction show why we want ordered pairs, for $x * y$ and $y * x$ may be distinct. As with any function, a binary operation is well-defined; when one says this explicitly, it is usually called the ***law of substitution***:

$$\text{If } x = x' \text{ and } y = y', \text{ then } x * y = x' * y'.$$

Definition. A ***group*** is a set G equipped with a binary operation $*$ such that

(i) the ***associative law*** holds: for every $x, y, z \in G$,

$$x * (y * z) = (x * y) * z;$$

(ii) there is an element $e \in G$, called the ***identity***, with $e * x = x = x * e$ for all $x \in G$;

(iii) every $x \in G$ has an ***inverse***; there is $x' \in G$ with $x * x' = e = x' * x$.

By Theorem 1.49, the set S_X of all permutations of a set X, with composition as the operation and $1_X = (1)$ as the identity, is a group (the ***symmetric group*** on X). In Exercise 2.22 on page 61, the reader will see that some of the equations in the definition of group are redundant. This is a useful observation, for it is more efficient, when verifying that a set with an operation is actually a group, to check fewer equations.

We are now at the precise point when algebra becomes *abstract* algebra. In contrast to the concrete group S_n consisting of all the permutations of $\{1, 2, \ldots, n\}$, we have passed to groups whose elements are unspecified. Moreover, products of elements are not explicitly computable but are, instead, merely subject to certain rules. It will be seen that this approach is quite fruitful, for theorems now apply to many different groups, and it is more efficient to prove theorems once for all instead of proving them anew for each group encountered. In addition to this obvious economy, it is often simpler to work with the "abstract" viewpoint even when dealing with a particular concrete group. For example, we will see that certain properties of S_n are simpler to treat without recognizing that the elements in question are permutations (see Example 2.26).

Definition. A group G is called ***abelian***[3] if it satisfies the ***commutative law***:

$$x * y = y * x$$

holds for every $x, y \in G$.

The groups S_n, for $n \geq 3$, are not abelian because $(1\ 2)$ and $(1\ 3)$ are elements of S_n that do not commute: $(1\ 2)(1\ 3) = (1\ 3\ 2)$ and $(1\ 3)(1\ 2) = (1\ 2\ 3)$.

Lemma 2.16. *Let G be a group.*

(i) *The **cancellation laws** hold: If either $x * a = x * b$ or $a * x = b * x$, then $a = b$.*

(ii) *The element e is the unique element in G with $e * x = x = x * e$ for all $x \in G$.*

(iii) *Each $x \in G$ has a unique inverse: There is only one element $x' \in G$ with $x * x' = e = x' * x$ (henceforth, this element will be denoted by x^{-1}).*

(iv) *$(x^{-1})^{-1} = x$ for all $x \in G$.*

Proof. (i) Choose x' with $x' * x = e = x * x'$; then

$$\begin{aligned} a = e * a = (x' * x) * a &= x' * (x * a) \\ &= x' * (x * b) = (x' * x) * b = e * b = b. \end{aligned}$$

A similar proof works when x is on the right.

(ii) Let $e_0 \in G$ satisfy $e_0 * x = x = x * e_0$ for all $x \in G$. In particular, setting $x = e$ in the second equation gives $e = e * e_0$; on the other hand, the defining property of e gives $e * e_0 = e_0$, so that $e = e_0$.

[3] The reason why commutative groups are called *abelian* can be found on page 236.

(iii) Assume that $x'' \in G$ satisfies $x * x'' = e = x'' * x$. Multiply the equation $e = x * x'$ on the left by x'' to obtain

$$x'' = x'' * e = x'' * (x * x') = (x'' * x) * x' = e * x' = x'.$$

(iv) By definition, $(x^{-1})^{-1} * x^{-1} = e = x^{-1} * (x^{-1})^{-1}$. But $x * x^{-1} = e = x^{-1} * x$, so that $(x^{-1})^{-1} = x$, by (iii). •

From now on, we will usually denote the product $x * y$ in a group by xy (we have already abbreviated $\alpha \circ \beta$ to $\alpha\beta$ in symmetric groups), and we will denote the identity by 1 instead of by e. When a group is abelian, however, we will often use the ***additive notation*** $x + y$; in this case, we will denote the identity by 0, and we will denote the inverse of an element x by $-x$ instead of by x^{-1}.

Example 2.17.

(i) The set $\mathbb{Q}^\times$ of all nonzero rationals is an abelian group, where $*$ is ordinary multiplication, the number 1 is the identity, and the inverse of $r \in \mathbb{Q}^\times$ is $1/r$. Similarly, $\mathbb{R}^\times$ and $\mathbb{C}^\times$ are multiplicative abelian groups.

Note that the set $\mathbb{Z}^\times$ of all nonzero integers is not a multiplicative group, for none of its elements (aside from ± 1) has a multiplicative inverse which is an integer.

(ii) The set $\mathbb{Z}$ of all integers is an additive abelian group with $a * b = a + b$, with identity $e = 0$, and with the inverse of an integer n being $-n$. Similarly, we can see that $\mathbb{Q}$, $\mathbb{R}$, and $\mathbb{C}$ are additive abelian groups.

(iii) The ***circle group***,

$$S^1 = \{z \in \mathbb{C} : |z| = 1\},$$

is the group whose operation is multiplication of complex numbers; this is an operation because the product of complex numbers of modulus 1 also has modulus 1, by Corollary 1.31. Complex multiplication is associative, the identity is 1 (which has modulus 1), and the inverse of any complex number of modulus 1 is its complex conjugate, which also has modulus 1. Therefore, S^1 is a group.

(iv) For any positive integer n, let

$$\mu_n = \left\{\zeta^k : 0 \le k < n\right\}$$

be the set of all the nth roots of unity, where

$$\zeta = e^{2\pi i/n} = \cos\left(\tfrac{2\pi}{n}\right) + i \sin\left(\tfrac{2\pi}{n}\right).$$

The reader may use De Moivre's theorem to see that μ_n is a group with operation multiplication of complex numbers; moreover, the inverse of any nth root of unity is its complex conjugate, which is also an nth root of unity.

(v) The plane $\mathbb{R} \times \mathbb{R}$ is a group with operation vector addition; that is, if $\alpha = (x, y)$ and $\alpha' = (x', y')$, then $\alpha + \alpha' = (x + x', y + y')$. The identity is the origin $O = (0, 0)$, and the inverse of (x, y) is $(-x, -y)$. ◄

Example 2.18.
Let X be a set. If U and V are subsets of X, define

$$U - V = \{x \in U : x \notin V\}.$$

The ***Boolean group*** $\mathcal{B}(X)$ [named after the logician G. Boole (1815–1864)] is the family of all the subsets of X equipped with addition given by ***symmetric difference*** $A + B$, where

$$A + B = (A - B) \cup (B - A);$$

symmetric difference is pictured in Figure 2.2.

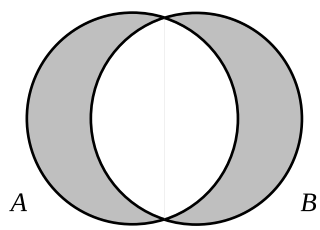


Figure 2.2

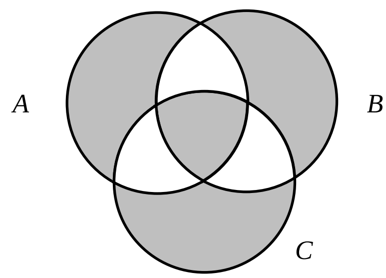


Figure 2.3

It is plain that $A + B = B + A$, so that symmetric difference is commutative. The identity is $\varnothing$, the empty set, and the inverse of A is A itself, for $A + A = \varnothing$. The reader may verify associativity by showing that both $(A + B) + C$ and $A + (B + C)$ are described by Figure 2.3. ◄

Example 2.19.
An $n \times n$ matrix A with real entries is called ***nonsingular*** if it has an inverse; that is, there is a matrix B with $AB = I = BA$, where $I = [\delta_{ij}]$ (δ_{ij} is the Kronecker delta) is the $n \times n$ identity matrix. Since $(AB)^{-1} = B^{-1}A^{-1}$, the product of nonsingular matrices is itself nonsingular. The set $\mathrm{GL}(n, \mathbb{R})$ of all $n \times n$ nonsingular matrices having real entries, with binary operation matrix multiplication, is a (nonabelian) group, called the ***general linear group***. [The proof of associativity is routine, though tedious; a "clean" proof of associativity can be given (Corollary 3.99) once the relation between matrices and linear transformations is known.] ◄

A binary operation allows us to multiply two elements at a time; how do we multiply three elements? There is a choice. Given the expression $2 \times 3 \times 4$, for example, we can first multiply $2 \times 3 = 6$ and then multiply $6 \times 4 = 24$; or, we can first multiply $3 \times 4 = 12$ and then multiply $2 \times 12 = 24$; of course, the answers agree, for multiplication of numbers is associative. Thus, if an operation is associative, the expression abc is not ambiguous. Not all operations are associative, however. For example, subtraction is not associative: if $c \neq 0$, then

$$a - (b - c) \neq (a - b) - c,$$

and so the notation $a - b - c$ is ambiguous. The cross product of two vectors in $\mathbb{R}^3$ is another example of a nonassociative operation.

Definition. If G is a group and if $a \in G$, define the ***powers***[4] a^n, for $n \geq 1$, inductively:

$$a^1 = a \quad \text{and} \quad a^{n+1} = aa^n.$$

Define $a^0 = 1$ and, if n is a positive integer, define

$$a^{-n} = (a^{-1})^n.$$

The reader expects that $(a^{-1})^n = (a^n)^{-1}$; this is a special case of the equation in Exercise 2.17 on page 61, but this is not so obvious to prove at this stage. For example, showing that $a^{-2}a^2 = 1$ amounts to doing the cancellation in the expression $(a^{-1}a^{-1})(aa)$; but associativity is given to us only for products having three, not four, factors.

Let us return to powers. The first and second powers are fine: $a^1 = a$ and $a^2 = aa$. There are two possible cubes: We have defined $a^3 = aa^2 = a(aa)$, but there is another reasonable contender: $(aa)a = a^2a$. If we assume associativity, then these are equal:

$$a^3 = aa^2 = a(aa) = (aa)a = a^2a.$$

There are several possible products of a with itself four times; assuming that the operation is associative, is it obvious that $a^4 = a^3a = a^2a^2$? And what about higher powers?

Define an ***expression*** $a_1a_2 \cdots a_n$ to be an n-tuple in $G \times \cdots \times G$ (n factors). An expression yields many elements of G by the following procedure. Choose two adjacent a's, multiply them, and obtain an expression with $n - 1$ factors: The new product just formed and $n - 2$ original factors. In this shorter new expression, choose two adjacent factors (either an original pair or an original one together with the new product from the first step) and multiply them. Repeat this procedure until there is an expression with only two

[4]The terminology x square and x cube for x^2 and x^3 is, of course, geometric in origin. Usage of the word *power* in this context arises from a mistranslation of the Greek *dunamis* (from which dynamo derives) used by Euclid. *Power* was the standard European rendition of *dunamis*; for example, the first English translation of Euclid, in 1570, by H. Billingsley, renders a sentence of Euclid as, "The power of a line is the square of the same line." However, contemporaries of Euclid (e.g., Aristotle and Plato) often used *dunamis* to mean amplification, and this seems to be a more appropriate translation, for Euclid was probably thinking of a one-dimensional line sweeping out a two-dimensional square. (I thank Donna Shalev for informing me of the classical usage of *dunamis*.)

factors; multiply them and obtain an element of G; call this an ***ultimate product*** derived from the expression. For example, consider the expression $abcd$. We may first multiply ab, obtaining $(ab)cd$, an expression with three factors, namely, ab, c, d. We may now choose either the pair c, d or the pair ab, c; in either case, multiply these, obtaining expressions with two factors: $(ab)(cd)$ having factors ab and cd or $((ab)c)d$ having factors $(ab)c$ and d. The two factors in either of these last expressions can now be multiplied to give an ultimate product from $abcd$. Other ultimate products derived from the expression $abcd$ arise by multiplying bc or cd as the first step. It is not obvious whether the ultimate products derived from a given expression are all equal.

Definition. An expression $a_1 a_2 \cdots a_n$ ***needs no parentheses*** if all the ultimate products it yields are equal; that is, no matter what choices are made of adjacent factors to multiply, all the resulting products in G are equal.

Theorem 2.20 (Generalized Associativity). *If G is a group and $a_1, a_2, \ldots, a_n \in G$, then the expression $a_1 a_2 \cdots a_n$ needs no parentheses.*

Remark. This result holds in greater generality, for neither the identity element nor inverses will be used in the proof. ◀

Proof. The proof is by (the second form of) induction. The base step $n = 3$ follows from associativity. For the inductive step, consider two ultimate products U and V obtained from an expression $a_1 a_2 \cdots a_n$ after two series of choices:

$$(a_1 \cdots a_i)(a_{i+1} \cdots a_n) \quad \text{and} \quad (a_1 \cdots a_j)(a_{j+1} \cdots a_n);$$

the parentheses indicate the last two factors which multiply to give U and V; there are many parentheses inside each of these shorter expressions. We may assume that $i \le j$. Since each of the four expressions in parentheses has fewer than n factors, the inductive hypothesis says that each needs no parentheses. It follows that $U = V$ if $i = j$. If $i < j$, then the inductive hypothesis allows the first expression to be rewritten

$$U = (a_1 \cdots a_i)\big([a_{i+1} \cdots a_j][a_{j+1} \cdots a_n]\big)$$

and the second to be rewritten

$$V = \big([a_1 \cdots a_i][a_{i+1} \cdots a_j]\big)(a_{j+1} \cdots a_n),$$

where each of the expressions $a_1 \cdots a_i$, $a_{i+1} \cdots a_j$, and $a_{j+1} \cdots a_n$ needs no parentheses. Thus, these expressions yield unique elements A, B, and C of G, respectively. The first expression yields $A(BC)$, the second yields $(AB)C$, and these two expressions give the same element of G, by associativity. •

Corollary 2.21. *If G is a group and $a, b \in G$, then*

$$(ab)^{-1} = b^{-1}a^{-1}.$$

Proof. By Lemma 2.16(iii), it suffices to prove that $(ab)(b^{-1}a^{-1}) = 1 = (b^{-1}a^{-1})(ab)$. Using generalized associativity,

$$(ab)(b^{-1}a^{-1}) = [a(bb^{-1})]a^{-1} = (a1)a^{-1} = aa^{-1} = 1.$$

A similar argument proves the other equation. •

Corollary 2.22. *If G is a group, if $a \in G$, and if $m, n \geq 1$, then*

$$a^{m+n} = a^m a^n \quad \textit{and} \quad (a^m)^n = a^{mn}.$$

Proof. In the first instance, both elements arise from the expression having $m + n$ factors each equal to a; in the second instance, both elements arise from the expression having mn factors each equal to a. •

It follows that any two powers of an element a in a group commute:

$$a^m a^n = a^{m+n} = a^{n+m} = a^n a^m.$$

Proposition 2.23 (Laws of Exponents). *Let G be a group, let $a, b \in G$, and let m and n be* (*not necessarily positive*) *integers.*

(i) *If a and b commute, then $(ab)^n = a^n b^n$.*

(ii) $(a^n)^m = a^{mn}$.

(iii) $a^m a^n = a^{m+n}$.

Sketch of Proof. The proofs, while routine, are lengthy double inductions. •

The notation a^n is the natural way to denote $a * a * \cdots * a$, where a appears n times. However, if the operation is +, then it is more natural to denote $a + a + \cdots + a$ by na. Let G be a group written additively; if $a, b \in G$ and m and n are (not necessarily positive) integers, then Proposition 2.23 is usually rewritten:

(i) $n(a + b) = na + nb$

(ii) $m(na) = (mn)a$

(iii) $ma + na = (m + n)a$

Definition. Let G be a group and let $a \in G$. If $a^k = 1$ for some $k \geq 1$, then the smallest such exponent $k \geq 1$ is called the ***order*** of a; if no such power exists, then one says that a has ***infinite order***.

The additive group of integers, $\mathbb{Z}$, is a group, and 3 is an element in it having infinite order (because $3 + 3 + \cdots + 3$ is never 0).

In any group G, the identity has order 1, and it is the only element of order 1; an element has order 2 if and only if it is equal to its own inverse.

The definition of order says that if x has order n and $x^m = 1$ for some positive integer m, then $n \leq m$. The next theorem says that n must be a divisor of m.

Theorem 2.24. *If $a \in G$ is an element of order n, then $a^m = 1$ if and only if $n \mid m$.*

Proof. Assume that $a^m = 1$. The division algorithm provides integers q and r with $m = nq + r$, where $0 \leq r < n$. It follows that $a^r = a^{m-nq} = a^m a^{-nq} = 1$. If $r > 0$, then we contradict n being the smallest positive integer with $a^n = 1$. Hence, $r = 0$ and $n \mid m$. Conversely, if $m = nk$, then $a^m = a^{nk} = (a^n)^k = 1^k = 1$. •

What is the order of a permutation in S_n?

Proposition 2.25. *Let $\alpha \in S_n$.*

(i) *If α is an r-cycle, then α has order r.*

(ii) *If $\alpha = \beta_1 \cdots \beta_t$ is a product of disjoint r_i-cycles β_i, then α has order* $\operatorname{lcm}\{r_1, \ldots, r_t\}$.

(iii) *If p is a prime, then α has order p if and only if it is a p-cycle or a product of disjoint p-cycles.*

Proof. (i) This is Exercise 2.5 on page 50.

(ii) Each β_i has order r_i, by (i). Suppose that $\alpha^M = (1)$. Since the β_i commute, $(1) = \alpha^M = (\beta_1 \cdots \beta_t)^M = \beta_1^M \cdots \beta_t^M$. By Exercise 2.11 on page 50, disjointness of the β's implies that $\beta_i^M = (1)$ for each i, so that Theorem 2.24 gives $r_i \mid M$ for all i; that is, M is a common multiple of $r_1, \ldots, r_t$. On the other hand, if $m = \operatorname{lcm}\{r_1, \ldots, r_t\}$, then it is easy to see that $\alpha^m = (1)$. Therefore, α has order m.

(iii) Write α as a product of disjoint cycles and use (ii). •

For example, a permutation in S_n has order 2 if and only if it is a transposition or a product of disjoint transpositions.

Example 2.26.
Suppose a deck of cards is shuffled, so that the order of the cards has changed from $1, 2, 3, 4, \ldots, 52$ to $2, 1, 4, 3, \ldots, 52, 51$. If we shuffle again in the same way, then the cards return to their original order. But a similar thing happens for any permutation α of the 52 cards: If one repeats α sufficiently often, the deck is eventually restored to its original order. One way to see this uses our knowledge of permutations. Write α as a product

of disjoint cycles, say, $\alpha = \beta_1\beta_2 \cdots \beta_t$, where β_i is an r_i-cycle. By Proposition 2.25, α has order k, where k is the least common multiple of the r_i. Therefore, $\alpha^k = (1)$.

Here is a more general result with a simpler proof (abstract algebra can be easier than algebra): If G is a finite group and $a \in G$, then $a^k = 1$ for some $k \geq 1$. Consider the subset $\{1, a, a^2, \ldots, a^n, \ldots\}$. Since G is finite, there must be a repetition occurring on this infinite list: There are integers $m > n$ with $a^m = a^n$, and hence $1 = a^m a^{-n} = a^{m-n}$. We have shown that there is some positive power of a equal to 1. [Our original argument that $\alpha^k = (1)$ for a permutation α of 52 cards is not worthless, because it gives an algorithm computing k.] ◄

Let us state what we have just proved in Example 2.26.

Proposition 2.27. *If G is a finite group, then every $x \in G$ has finite order.*

Table 2.3 augments the table in Example 2.5(ii).

Cycle Structure	Number	Order	Parity
(1)	1	1	Even
(1 2)	10	2	Odd
(1 2 3)	20	3	Even
(1 2 3 4)	30	4	Odd
(1 2 3 4 5)	24	5	Even
(1 2)(3 4 5)	20	6	Odd
(1 2)(3 4)	15	2	Even
	120		

Table 2.3. Permutations in S_5

Here are some geometric examples of groups.

Definition. A ***motion*** is a distance preserving bijection $\varphi \colon \mathbb{R}^2 \to \mathbb{R}^2$ [it can be shown that φ is a linear transformation if $\varphi(0) = 0$]. If π is a polygon in the plane, then its ***symmetry group*** $\Sigma(\pi)$ consists of all the motions φ for which $\varphi(\pi) = \pi$. The elements of $\Sigma(\pi)$ are called ***symmetries*** of π.

Example 2.28.
(i) Let π_4 be a square having sides of length 1 and vertices $\{v_1, v_2, v_3, v_4\}$; draw π_4 in the plane so that its center is at the origin O and its sides are parallel to the axes. It can be shown that every $\varphi \in \Sigma(\pi_4)$ permutes the vertices; indeed, a symmetry φ of π_4 is determined by $\{\varphi(v_i) : 1 \leq i \leq 4\}$, and so there are at most $24 = 4!$ possible symmetries. Not every permutation in S_4 arises from a symmetry of π_4, however. If v_i and v_j are adjacent, then $\|v_i - v_j\| = 1$, but $\|v_1 - v_3\| = \sqrt{2} = \|v_2 - v_4\|$; it follows that φ must preserve adjacency (for motions preserve distance). The reader may now check that there are only eight symmetries of π_4. Aside from the identity and the three rotations about O

by 90°, 180°, and 270°, there are four reflections, respectively, in the lines v_1v_3, v_2v_4, the x-axis, and the y-axis (for a generalization to come, note that the y-axis is Om_1, where m_1 is the midpoint of v_1v_2, and the x-axis is Om_2, where m_2 is the midpoint of v_2v_3). The group $\Sigma(\pi_4)$ is called the ***dihedral group***[5] with 8 elements, and it is denoted by D_8.

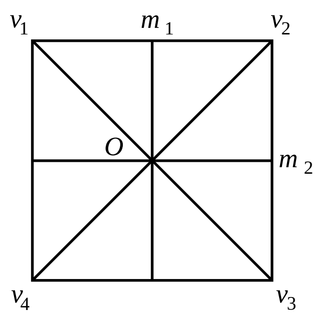


Figure 2.4

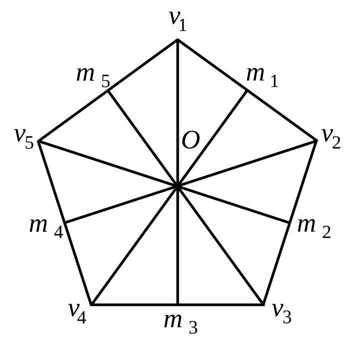


Figure 2.5

(ii) The symmetry group $\Sigma(\pi_5)$ of a regular pentagon π_5 with vertices $v_1, \ldots, v_5$ and center O has 10 elements: the rotations about the origin of $(72j)°$, where $0 \le j \le 4$, as well as the reflections in the lines Ov_k for $1 \le k \le 5$. The symmetry group $\Sigma(\pi_5)$ is called the ***dihedral group*** with 10 elements, and it is denoted by D_{10}. ◄

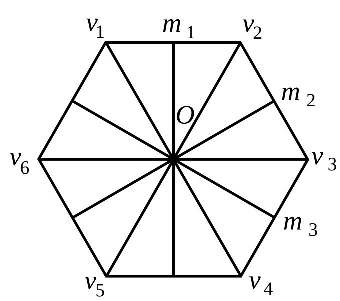


Figure 2.6

[5]F. Klein was investigating those finite groups occurring as subgroups of the group of motions of $\mathbb{R}^3$. Some of these occur as symmetry groups of regular polyhedra (from the Greek *poly* meaning "many" and *hedron* meaning "two-dimensional side"). He invented a degenerate polyhedron that he called a *dihedron*, from the Greek words *di* meaning "two" and *hedron*, which consists of two congruent regular polygons of zero thickness pasted together. The symmetry group of a dihedron is thus called a *dihedral group*. For our purposes, it is more natural to describe these groups as in the text.

Definition. If π_n is a regular polygon with n vertices $v_1, v_2, \ldots, v_n$ and center O, then the symmetry group $\Sigma(\pi_n)$ is called the ***dihedral group*** with $2n$ elements, and it is denoted[6] by D_{2n}.

The dihedral group D_{2n} contains the n rotations ρ^j about the center by $(360j/n)^\circ$, where $0 \le j \le n-1$. The description of the other n elements depends on the parity of n. If n is odd (as in the case of the pentagon; see Figure 2.5), then the other n symmetries are reflections in the distinct lines Ov_i, for $i = 1, 2, \ldots, n$. If $n = 2q$ is even (see the square in Figure 2.4 or the regular hexagon in Figure 2.6), then each line Ov_i coincides with the line Ov_{q+i}, giving only q such reflections; the remaining q symmetries are reflections in the lines Om_i for $i = 1, 2, \ldots, q$, where m_i is the midpoint of the edge $v_i v_{i+1}$. For example, the six lines of symmetry of π_6 are Ov_1, Ov_2, and Ov_3, and Om_1, Om_2, and Om_3.

EXERCISES

2.17 If $a_1, a_2, \ldots, a_{t-1}, a_t$ are elements in a group G, prove that

$$(a_1 a_2 \cdots a_{t-1} a_t)^{-1} = a_t^{-1} a_{t-1}^{-1} \cdots a_2^{-1} a_1^{-1}.$$

2.18 Assume that G is a set with an associative binary operation. Prove that $(ab)(cd) = a[(bc)d]$ without using generalized associativity.

2.19 (i) Compute the order, inverse, and parity of

$$\alpha = (1\ 2)(4\ 3)(1\ 3\ 5\ 4\ 2)(1\ 5)(1\ 3)(2\ 3).$$

(ii) What are the respective orders of the permutations in Exercises 2.1 on page 49 and 2.8 on page 50?

2.20 (i) How many elements of order 2 are there in S_5 and in S_6?

(ii) How many elements of order 2 are there in S_n?

Hint. You may express your answer as a sum.

2.21 If G is a group, prove that the only element $g \in G$ with $g^2 = g$ is 1.

2.22 This exercise gives a shorter list of axioms defining a group. Let H be a set containing an element e, and assume that there is an associative binary operation $*$ on H satisfying the following properties:

1. $e * x = x$ for all $x \in H$;

2. for every $x \in H$, there is $x' \in H$ with $x' * x = e$.

(i) Prove that if $h \in H$ satisfies $h * h = h$, then $h = e$.

Hint. If $h' * h = e$, evaluate $h' * h * h$ in two ways.

(ii) For all $x \in H$, prove that $x * x' = e$.

Hint. Consider $(x * x')^2$.

(iii) For all $x \in H$, prove that $x * e = x$.

Hint. Evaluate $x * x' * x$ in two ways.

[6] Some authors denote D_{2n} by D_n.

(iv) Prove that if $e' \in H$ satisfies $e' * x = x$ for all $x \in H$, then $e' = e$.

Hint. Show that $(e')^2 = e'$.

(v) Let $x \in H$. Prove that if $x'' \in H$ satisfies $x'' * x = e$, then $x'' = x'$.

Hint. Evaluate $x' * x * x''$ in two ways.

(vi) Prove that H is a group.

2.23 Let y be a group element of order m; if $m = pt$ for some prime p, prove that y^t has order p.
Hint. Clearly, $(y^t)^p = 1$. Use Theorem 2.24 to show that no smaller power of y^t is equal to 1.

2.24 Let G be a group and let $a \in G$ have order k. If p is a prime divisor of k, and if there is $x \in G$ with $x^p = a$, prove that x has order pk.

2.25 Let $G = \mathrm{GL}(2, \mathbb{Q})$, and let

$$A = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} 0 & 1 \\ -1 & 1 \end{bmatrix}.$$

Show that $A^4 = I = B^6$, but that $(AB)^n \neq I$ for all $n > 0$, where $I = \left[\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right]$ is the 2×2 identity matrix. Conclude that AB can have infinite order even though both factors A and B have finite order (this cannot happen in a finite group).

2.26 If G is a group in which $x^2 = 1$ for every $x \in G$, prove that G must be abelian. [The Boolean groups $\mathcal{B}(X)$ of Example 2.18 are such groups.]

2.27 If G is a group with an even number of elements, prove that the number of elements in G of order 2 is odd. In particular, G must contain an element of order 2.
Hint. Pair each element with its inverse.

2.28 What is the largest order of an element in S_n, where $n = 1, 2, \ldots, 10$? (We remark that no general formula is known for arbitrary n, although, in 1903, E. Landau found the asymptotic behavior.)

2.4 Lagrange's Theorem

A *subgroup* H of a group G is a group contained in G so that if $h, h' \in H$, then the product hh' in H is the same as the product hh' in G. The formal definition of subgroup, however, is more convenient to use.

Definition. A subset H of a group G is a ***subgroup*** if

(i) $1 \in H$;

(ii) if $x, y \in H$, then $xy \in H$;

(iii) if $x \in H$, then $x^{-1} \in H$.

If H is a subgroup of G, we write $H \leq G$; if H is a proper subgroup of G, that is, $H \neq G$, then we write $H < G$.

Observe that $\{1\}$ and G are always subgroups of a group G, where $\{1\}$ denotes the subset consisting of the single element 1. More interesting examples will be given soon. A subgroup $H \neq G$ is called a ***proper subgroup***.

Let us see that every subgroup $H \leq G$ is itself a group. Property (ii) shows that H is ***closed***; that is, H has a binary operation. Associativity $(xy)z = x(yz)$ holds for all $x, y, z \in G$, and so this equation holds, in particular, for all $x, y, z \in H$. Finally, (i) gives the identity, and (iii) gives inverses.

It is easier to check that a subset H of a group G is a subgroup (and hence that it is a group in its own right) than to verify the group axioms for H: Associativity is inherited from the operation on G and hence it need not be verified again.

Example 2.29.
(i) The four permutations

$$\mathbf{V} = \big\{(1), (1\ 2)(3\ 4), (1\ 3)(2\ 4), (1\ 4)(2\ 3)\big\}$$

form a group, because $\mathbf{V}$ is a subgroup of S_4 : $(1) \in \mathbf{V}$; $\alpha^2 = (1)$ for each $\alpha \in \mathbf{V}$, and so $\alpha^{-1} = \alpha \in \mathbf{V}$; the product of any two distinct permutations in $\mathbf{V} - \{(1)\}$ is the third one. The group $\mathbf{V}$ is called the ***four-group*** ($\mathbf{V}$ abbreviates the original German term *Vierergruppe*).

Consider what verifying associativity $a(bc) = (ab)c$ would involve: There are 4 choices for each of a, b, and c, and so there are $4^3 = 64$ equations to be checked. Plainly, the best way to prove that $\mathbf{V}$ is a group is to show that it is a subgroup of S_4.

(ii) If $\mathbb{R}^2$ is the plane considered as an (additive) abelian group, then any line L through the origin is a subgroup. The easiest way to see this is to choose a point $(a, b) \neq (0, 0)$ on L and then note that L consists of all the scalar multiples (ra, rb). The reader may now verify that the axioms in the definition of subgroup do hold for L. ◂

We can shorten the list of items needed to verify that a subset is, in fact, a subgroup.

Proposition 2.30. *A subset H of a group G is a subgroup if and only if H is nonempty and, whenever $x, y \in H$, then $xy^{-1} \in H$.*

Proof. Necessity is clear. For sufficiency, take $x \in H$ (which exists because $H \neq \varnothing$); by hypothesis, $1 = xx^{-1} \in H$. If $y \in H$, then $y^{-1} = 1y^{-1} \in H$ and, if $x, y \in H$, then $xy = x(y^{-1})^{-1} \in H$. •

Of course, the simplest way to check that a candidate H for a subgroup is nonempty is to check whether $1 \in H$.

Note that if the operation in G is addition, then the condition in the proposition is that H is a nonempty subset such that $x, y \in H$ implies $x - y \in H$.

Proposition 2.31. *A nonempty subset H of a finite group G is a subgroup if and only if H is closed; that is, if $a, b \in H$, then $ab \in H$. In particular, a nonempty subset of S_n is a subgroup if and only if it is closed.*

Sketch of Proof. Since G is finite, Proposition 2.27 says that each $x \in G$ has finite order. Hence, if $x^n = 1$, then $1 \in H$ and $x^{-1} = x^{n-1} \in H$. •

This last proposition can be false when G is an infinite group. For example, let G be the additive group $\mathbb{Z}$; the subset $H = \mathbb{N}$ is closed, but it is not a subgroup of $\mathbb{Z}$.

For Galois, in 1830, a group was just a subset H of S_n that is closed under composition; that is, if $\alpha, \beta \in H$, then $\alpha\beta \in H$. A. Cayley, in 1854, was the first to define an abstract group, mentioning associativity, inverses, and identity explicitly. He then proved (see Cayley's theorem) that every abstract group with n elements is, essentially, a subgroup of S_n (the notion of isomorphism, introduced in the next section, will enable us to state this more precisely).

Example 2.32.
The subset A_n of S_n, consisting of all the even permutations, is a subgroup because it is closed under multiplication: even $\circ$ even $=$ even. This subgroup $A_n \leq S_n$ is called the ***alternating***[7] ***group*** on n letters. ◂

Definition. If G is a group and $a \in G$, write

$$\langle a \rangle = \{a^n : n \in \mathbb{Z}\} = \{\text{all powers of } a\};$$

$\langle a \rangle$ is called the ***cyclic subgroup*** of G ***generated*** by a. A group G is called ***cyclic*** if there exists $a \in G$ with $G = \langle a \rangle$, in which case a is called a ***generator*** of G.

It is easy to see that $\langle a \rangle$ is, in fact, a subgroup: $1 = a^0 \in \langle a \rangle$; $a^n a^m = a^{n+m} \in \langle a \rangle$; $a^{-1} \in \langle a \rangle$. Example 2.17(iv) shows, for every $n \geq 1$, that the multiplicative group μ_n of all nth roots of unity is a cyclic group with the primitive nth root of unity $\zeta = e^{2\pi i/n}$ as a generator.

No doubt, the reader has seen the example of the integers modulo m in an earlier course. We merely recall the definition. Given $m \geq 0$ and $a \in \mathbb{Z}$, the *congruence class* $[a]$ of a mod m was defined on page 34:

$$\begin{aligned}[a] &= \{b \in \mathbb{Z} : b \equiv a \bmod m\} \\ &= \{a + km : k \in \mathbb{Z}\} \\ &= \{\ldots, a - 2m, a - m, a, a + m, a + 2m, \ldots\}.\end{aligned}$$

[7]The *alternating group* first arose in studying polynomials. If

$$f(x) = (x - u_1)(x - u_2) \cdots (x - u_n),$$

then the number $D = \prod_{i<j}(u_i - u_j)$ can change sign when the roots are permuted: If α is a permutation of $\{u_1, u_2, \ldots, u_n\}$, then it is easy to see that $\prod_{i<j}[\alpha(u_i) - \alpha(u_j)] = \pm D$. Thus, the sign of the product alternates as various permutations α are applied to its factors. The sign does not change for those α in the alternating group, and this last fact can be used to give another proof of Theorem 2.13(ii).

Definition. The ***integers mod m***, denoted[8] by $\mathbb{I}_m$, is the family of all congruence classes mod m.

Recall that $[a] = [b]$ in $\mathbb{I}_m$ if and only if $a \equiv b \bmod m$. In particular, $[a] = [0]$ in $\mathbb{I}_m$ if and only if $a \equiv 0 \bmod m$; that is, $[a] = [0]$ in $\mathbb{I}_m$ if and only if m is a divisor of a. The definition of congruence mod m makes sense for all $m \geq 0$, but the cases $m = 0$ and $m = 1$ are not very interesting: $a \equiv b \bmod 0$ means $0 \mid (a - b)$, which says that $a = b$; $a \equiv b \bmod 1$ means $1 \mid (a - b)$, which says that a and b are always congruent; that is, there is only one congruence class mod 1. Recall Proposition 1.19, which we now rewrite in the bracket notation.

Proposition 1.19. *Let $m \geq 2$ be a fixed integer.*

(i) *If $a \in \mathbb{Z}$, then $[a] = [r]$ for some r with $0 \leq r < m$.*

(ii) *If $0 \leq r' < r < m$, then $[r'] \neq [r]$.*

(iii) *$\mathbb{I}_m$ has exactly m elements, namely, $[0], [1], \ldots, [m-1]$.*

For every $m \geq 2$, $\mathbb{I}_m$ is an (additive) cyclic group, where

$$[a] + [b] = [a + b];$$

the identity is $[0]$, the inverse of $[a]$ is $[-a]$, and a generator is $[1]$. Part (iii) shows that $\mathbb{I}_m$ has order m.

A cyclic group can have several different generators. For example, $\langle a \rangle = \langle a^{-1} \rangle$.

Theorem 2.33.

(i) *If $G = \langle a \rangle$ is a cyclic group of order n, then a^k is a generator of G if and only if $(k, n) = 1$.*

(ii) *If G is a cyclic group of order n and* $\text{gen}(G) = \{\text{all generators of } G\}$, *then*

$$|\text{gen}(G)| = \phi(n),$$

where ϕ is the Euler ϕ-function.

Proof. (i) If a^k generates G, then $a \in \langle a^k \rangle$, so that $a = a^{kt}$ for some $t \in \mathbb{Z}$. Hence, $a^{kt-1} = 1$; by Theorem 2.24, $n \mid (kt - 1)$, so there is $v \in \mathbb{Z}$ with $nv = kt - 1$. Therefore, 1 is a linear combination of k and m, and so $(k, n) = 1$.

Conversely, if $(k, n) = 1$, then $nt + ku = 1$ for $t, u \in \mathbb{Z}$; hence

$$a = a^{nt+ku} = a^{nt}a^{ku} = a^{ku} \in \langle a^k \rangle.$$

Therefore, every power of a also lies in $\langle a^k \rangle$ and $G = \langle a^k \rangle$.

(ii) Proposition 1.38 says that $\phi(n) = |\{k \leq n : (k, n) = 1\}|$. The next proposition shows that $G = \{1, a, \ldots, a^{n-1}\}$, and so this result follows from part (i). •

[8]We introduce this new notation because there is no commonly agreed one; the most popular contenders are $\mathbb{Z}_m$ and $\mathbb{Z}/m\mathbb{Z}$. We have chosen $\mathbb{I}_m$ because I is the initial letter of *integers*. The usual notation $\mathbb{Z}$ for the integers (it is the initial letter of the German *Zahlen*) is almost universally accepted, and so a change from $\mathbb{Z}$ to $\mathbb{I}$ would be consistent but too distracting.

Proposition 2.34. *Let G be a finite group and let $a \in G$. Then the order of a is $|\langle a \rangle|$, the number of elements in $\langle a \rangle$.*

Proof. Since G is finite, there is an integer $k \geq 1$ with $1, a, a^2, \ldots, a^{k-1}$ consisting of k distinct elements, while $1, a, a^2, \ldots, a^k$ has a repetition; hence $a^k \in \{1, a, a^2, \ldots, a^{k-1}\}$; that is, $a^k = a^i$ for some i with $0 \leq i < k$. If $i \geq 1$, then $a^{k-i} = 1$, contradicting the original list having no repetitions. Therefore, $a^k = a^0 = 1$, and k is the order of a (being the smallest positive such k).

If $H = \{1, a, a^2, \ldots, a^{k-1}\}$, then $|H| = k$; it suffices to show that $H = \langle a \rangle$. Clearly, $H \subseteq \langle a \rangle$. For the reverse inclusion, take $a^i \in \langle a \rangle$. By the division algorithm, $i = qk + r$, where $0 \leq r < k$. Hence $a^i = a^{qk+r} = a^{qk}a^r = (a^k)^q a^r = a^r \in H$; this gives $\langle a \rangle \subseteq H$, and so $\langle a \rangle = H$. •

Definition. If G is a finite group, then the number of elements in G, denoted by $|G|$, is called the ***order*** of G.

The word *order* is used in two senses: the order of an *element* $a \in G$ and the order $|G|$ of a *group* G. Proposition 2.34 shows that the order of a group element a is equal to $|\langle a \rangle|$.

Proposition 2.35. *The intersection $\bigcap_{i \in I} H_i$ of any family of subgroups of a group G is again a subgroup of G. In particular, if H and K are subgroups of G, then $H \cap K$ is a subgroup of G.*

Sketch of Proof. This follows easily from the definitions. •

Corollary 2.36. *If X is a subset of a group G, then there is a subgroup $\langle X \rangle$ of G containing X that is* ***smallest*** *in the sense that $\langle X \rangle \leq H$ for every subgroup H of G that contains X.*

Proof. There exist subgroups of G that contain X; for example, G itself contains X. Define $\langle X \rangle = \bigcap_{X \subseteq H} H$, the intersection of all the subgroups H of G that contain X. By Proposition 2.35, $\langle X \rangle$ is a subgroup of G; of course, $\langle X \rangle$ contains X because every H contains X. Finally, if H is any subgroup containing X, then H is one of the subgroups whose intersection is $\langle X \rangle$; that is, $\langle X \rangle \leq H$. •

Note that there is no restriction on the subset X in the last corollary; in particular, $X = \varnothing$ is allowed. Since the empty set is a subset of every set, we have $\varnothing \subseteq H$ for every subgroup H of G. Thus, $\langle \varnothing \rangle$ is the intersection of *all* the subgroups of G; in particular, $\langle \varnothing \rangle \leq \{1\}$, and so $\langle \varnothing \rangle = \{1\}$.

Definition. If X is a subset of a group G, then $\langle X \rangle$ is called the ***subgroup generated by*** X.

If X is a nonempty subset of a group G, define a ***word***[9] on X to be an element $g \in G$ of the form $g = x_1^{e_1} \cdots x_n^{e_n}$, where $x_i \in X$ and $e_i = \pm 1$ for all i.

[9]This term will be modified a bit when we discuss free groups.

Proposition 2.37. *If X is a nonempty subset of a group G, then $\langle X \rangle$ is the set of all the words on X.*

Proof. We claim that $W(X)$, the set of all the words on X, is a subgroup. If $x \in X$, then $1 = xx^{-1} \in W(X)$; the product of two words on X is also a word on X; the inverse of a word on X is a word on X. It now follows that $\langle X \rangle \le W(X)$, for $W(X)$ obviously contains X (and $\langle X \rangle$ is the intersection of all the subgroups of G containing X). On the other hand, any subgroup of G containing X must also contain $W(X)$, and so $\langle X \rangle = W(X)$. •

Example 2.38.
(i) If $G = \langle a \rangle$ is a cyclic group with generator a, then G is generated by the subset $X = \{a\}$.

(ii) The dihedral group D_{2n}, the symmetry group of a regular n-gon, is generated by ρ, σ, where ρ is a rotation by $(360/n)°$ and σ is a reflection. Note that these generators satisfy the equations $\rho^n = 1$, $\sigma^2 = 1$, and $\sigma\rho\sigma = \rho^{-1}$. ◂

Perhaps the most fundamental fact about subgroups H of a finite group G is that their orders are constrained. Certainly, we have $|H| \le |G|$, but it turns out that $|H|$ must be a divisor of $|G|$. To prove this, we introduce the notion of coset.

Definition. If H is a subgroup of a group G and $a \in G$, then the ***coset*** aH is the subset aH of G, where

$$aH = \{ah : h \in H\}.$$

The cosets defined are often called ***left cosets***; there are also ***right cosets*** of H, namely, subsets of the form $Ha = \{ha : h \in H\}$. In general, left cosets and right cosets may be different, as we shall soon see.

If we use the $*$ notation for the operation in a group G, then we denote the coset aH by $a * H$, where

$$a * H = \{a * h : h \in H\}.$$

In particular, if the operation is addition, then the coset is denoted by

$$a + H = \{a + h : h \in H\}.$$

Of course, $a = a1 \in aH$. Cosets are usually not subgroups. For example, if $a \notin H$, then $1 \notin aH$ (otherwise $1 = ah$ for some $h \in H$, and this gives the contradiction $a = h^{-1} \in H$).

Example 2.39.
(i) Consider the plane $\mathbb{R}^2$ as an (additive) abelian group and let L be a line through the origin O (see Figure 2.7 on page 68); as in Example 2.29(ii), the line L is a subgroup of $\mathbb{R}^2$. If $\beta \in \mathbb{R}^2$, then the coset $\beta + L$ is the line L' containing β that is parallel to L, for if $r\alpha \in L$, then the parallelogram law gives $\beta + r\alpha \in L'$.

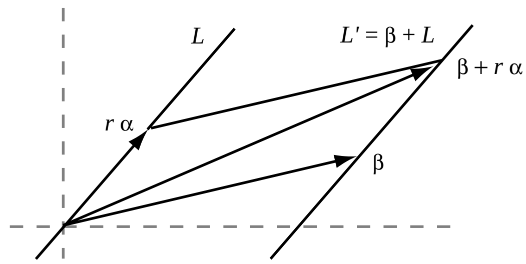


Figure 2.7

(ii) Let A be an $m \times n$ matrix with real entries, and let $A\mathbf{x} = \mathbf{b}$ be a *consistent* linear system of equations; that is, there is a column vector $\mathbf{s} \in \mathbb{R}^n$ with $A\mathbf{s} = \mathbf{b}$. The ***solution space*** $S = \{\mathbf{x} \in \mathbb{R}^n : A\mathbf{x} = \mathbf{0}\}$ of the homogeneous system $A\mathbf{x} = \mathbf{0}$ is an additive subgroup of $\mathbb{R}^n$, and the ***solution set*** $\{\mathbf{x} \in \mathbb{R}^n : A\mathbf{x} = \mathbf{b}\}$ of the original inhomogeneous system is the coset $\mathbf{s} + S$.

(iii) If $G = S_3$ and $H = \langle(1\ 2)\rangle$, there are exactly three left cosets of H, namely

$$\begin{aligned} H = \quad & \{(1), (1\ 2)\} \;=\; (1\ 2)H, \\ (1\ 3)H = & \{(1\ 3), (1\ 2\ 3)\} = (1\ 2\ 3)H, \\ (2\ 3)H = & \{(2\ 3), (1\ 3\ 2)\} = (1\ 3\ 2)H, \end{aligned}$$

each of which has size 2. Note that these cosets are also “parallel;” that is, distinct cosets are disjoint.

Consider the right cosets of $H = \langle(1\ 2)\rangle$ in S_3:

$$\begin{aligned} H = \quad & \{(1), (1\ 2)\} \;=\; H(1\ 2), \\ H(1\ 3) = & \{(1\ 3), (1\ 3\ 2)\} = H(1\ 3\ 2), \\ H(2\ 3) = & \{(2\ 3), (1\ 2\ 3)\} = H(1\ 2\ 3). \end{aligned}$$

Again, we see that there are exactly 3 (right) cosets, each of which has size 2. Note that these cosets are “parallel”; that is, distinct (right) cosets are disjoint. ◂

Lemma 2.40. *Let H be a subgroup of a group G, and let $a, b \in G$.*

(i) *$aH = bH$ if and only if $b^{-1}a \in H$. In particular, $aH = H$ if and only if $a \in H$.*

(ii) *If $aH \cap bH \neq \varnothing$, then $aH = bH$.*

(iii) *$|aH| = |H|$ for all $a \in G$.*

Remark. In Exercise 2.29 on page 72, it is shown that $Ha = Hb$ if and only if $ab^{-1} \in H$, and hence $Ha = H$ if and only if $a \in H$. ◀

Sketch of Proof. The first two statements follow from observing that the relation on G, defined by $a \equiv b$ if $b^{-1}a \in H$, is an equivalence relation whose equivalence classes are the cosets; it follows from Proposition 1.54 that the cosets of H partition G. The third statement is true because $h \mapsto ah$ is a bijection $H \to aH$. •

The next theorem is named after J. L. Lagrange, who saw, in 1770, that the order of certain subgroups of S_n are divisors of $n!$. The notion of group was invented by Galois 60 years afterward, and it was probably Galois who first proved the theorem in full.

Theorem 2.41 (Lagrange's Theorem). *If H is a subgroup of a finite group G, then $|H|$ is a divisor of $|G|$.*

Proof. Let $\{a_1H, a_2H, \ldots, a_tH\}$ be the family of all the distinct cosets of H in G. Then

$$G = a_1H \cup a_2H \cup \cdots \cup a_tH,$$

because each $g \in G$ lies in the coset gH, and $gH = a_iH$ for some i. Moreover, Lemma 2.40(ii) shows that the cosets partition G into pairwise disjoint subsets. It follows that

$$|G| = |a_1H| + |a_2H| + \cdots + |a_tH|.$$

But $|a_iH| = |H|$ for all i, by Lemma 2.40(iii), so that $|G| = t|H|$, as desired. •

Definition. The ***index*** of a subgroup H in G, denoted by $[G : H]$, is the number of left[10] cosets of H in G.

The index $[G : H]$ is the number t in the formula $|G| = t|H|$ in the proof of Lagrange's theorem, so that

$$|G| = [G : H]|H|;$$

this formula shows that the index $[G : H]$ is also a divisor of $|G|$; moreover,

$$[G : H] = |G|/|H|.$$

Example 2.42.
Recall that the dihedral group $D_{2n} = \Sigma(\pi_n)$, the group of symmetries of the regular n-gon π_n, has order $2n$ and it contains a cyclic subgroup of order n generated by a rotation ρ. The subgroup $\langle\rho\rangle$ has index $[D_{2n} : \langle\rho\rangle] = 2$. Thus, there are two cosets: $\langle\rho\rangle$ and $\sigma\,\langle\rho\rangle$, where σ is any reflection outside of $\langle\rho\rangle$. It follows that every element $\alpha \in D_{2n}$ has a factorization $\alpha = \sigma^i\rho^j$, where $i = 0, 1$ and $0 \le j < n$. ◀

[10]Exercise 2.37 on page 72 shows that the number of left cosets of a subgroup is equal to the number of its right cosets.

Corollary 2.43. *If G is a finite group and $a \in G$, then the order of a is a divisor of $|G|$.*

Proof. This follows at once from Proposition 2.34, for the order of a is $|\langle a \rangle|$. •

Corollary 2.44. *If G is a finite group, then $a^{|G|} = 1$ for all $a \in G$.*

Proof. If a has order d, then $|G| = dm$ for some integer m, by the previous corollary, and so $a^{|G|} = a^{dm} = (a^d)^m = 1$. •

Corollary 2.45. *If p is a prime, then every group G of order p is cyclic.*

Proof. If $a \in G$ and $a \neq 1$, then a has order $d > 1$, and d is a divisor of p. Since p is prime, $d = p$, and so $G = \langle a \rangle$. •

We have seen that $\mathbb{I}_m$, under addition, is a cyclic group of order m. Now multiplication $\mu : \mathbb{I}_m \times \mathbb{I}_m \to \mathbb{I}_m$, given by

$$[a][b] = [ab],$$

is also a binary operation on $\mathbb{I}_m$ (which is well-defined, by Proposition 1.20); it is associative, commutative, and [1] is an identity element. However, $\mathbb{I}_m$ is not a group under this operation because inverses may not exist; for example, [0] has no multiplicative inverse.

Proposition 2.46. *The set $U(\mathbb{I}_m)$, defined by*

$$U(\mathbb{I}_m) = \{\, [r] \in \mathbb{I}_m : (r, m) = 1\},$$

is a multiplicative group of order $\phi(m)$, where ϕ is the Euler ϕ-function. In particular, if p is a prime, then $U(\mathbb{I}_p) = \mathbb{I}_p^\times$, the nonzero elements of $\mathbb{I}_p$, is a multiplicative group of order $p - 1$.

Proof. By Exercise 1.14 on page 12, $(r, m) = 1 = (r', m)$ implies $(rr', m) = 1$; hence $U(\mathbb{I}_m)$ is closed under multiplication. We have already mentioned that multiplication is associative and that [1] is the identity. If $(a, m) = 1$, then $[a][x] = [1]$ can be solved for $[x]$ in $\mathbb{I}_m$. Now $(x, m) = 1$, for $rx + sm = 1$ for some integer s, and so Proposition 1.13 on page 5 gives $(x, m) = 1$; therefore, $[x] \in U(\mathbb{I}_m)$, and so each $[r] \in U(\mathbb{I}_m)$ has an inverse. Therefore, $U(\mathbb{I}_m)$ is a group; the definition of the Euler ϕ-function shows that $|U(\mathbb{I}_m)| = \phi(m)$.

The last statement follows from $\phi(p) = p - 1$ when p is a prime. •

In Chapter 3, we will prove, for every prime p, that $\mathbb{I}_p^\times$ is a cyclic group.

Here is a group-theoretic proof of Theorem 1.24, Fermat's theorem. Our earlier proof used binomial coefficients and the fact that $p \mid \binom{p}{r}$ for $0 < r < p$.

Corollary 2.47 (*Fermat*). *If p is a prime and $a \in \mathbb{Z}$, then*

$$a^p \equiv a \bmod p.$$

Proof. It suffices to show that $[a^p] = [a]$ in $\mathbb{I}_p$. If $[a] = [0]$, then $[a^p] = [a]^p = [0]^p = [0] = [a]$. If $[a] \neq [0]$, then $[a] \in \mathbb{I}_p^\times$, the multiplicative group of nonzero elements in $\mathbb{I}_p$. By Corollary 2.44 to Lagrange's theorem, $[a]^{p-1} = [1]$, because $|\mathbb{I}_p^\times| = p-1$. Multiplying by $[a]$ gives the desired result $[a^p] = [a]^p = [a]$. Therefore, $a^p \equiv a \bmod p$. •

We now give a generalization of Fermat's theorem due to Euler.

Theorem 2.48 (*Euler*). *If* $(r, m) = 1$, *then*

$$r^{\phi(m)} \equiv 1 \bmod m.$$

Proof. Since $|U(\mathbb{I}_m)| = \phi(m)$, Corollary 2.44 (essentially Lagrange's theorem) gives $[r]^{\phi(m)} = [1]$ for all $[r] \in U(\mathbb{I}_m)$. In congruence notation, this says that if $(r, m) = 1$, then $r^{\phi(m)} \equiv 1 \bmod m$. •

Example 2.49.
It is easy to see that

$$U(\mathbb{I}_8) = \big\{[1], [3], [5], [7]\big\}$$

is a group (resembling the four-group **V**) in which the square of each element is [1], while

$$U(\mathbb{I}_{10}) = \big\{[1], [3], [7], [9]\big\}$$

is a cyclic group of order 4 [after we introduce *isomorphisms* in the next section, we will say that $U(\mathbb{I}_8)$ is isomorphic to **V** and $U(\mathbb{I}_{10})$ is isomorphic to $\mathbb{I}_4$]. ◂

Theorem 2.50 (Wilson's Theorem). *An integer* p *is a prime if and only if*

$$(p-1)! \equiv -1 \bmod p.$$

Proof. Assume that p is a prime. If $a_1, a_2, \ldots, a_n$ is a list of all the elements of a finite abelian group G, then the product $a_1 a_2 \ldots a_n$ is the same as the product of all elements a with $a^2 = 1$, for any other element cancels against its inverse. Since p is prime, Exercise 1.37 on page 14 implies that $\mathbb{I}_p^\times$ has only one element of order 2, namely, $[-1]$. It follows that the product of all the elements in $\mathbb{I}_p^\times$, namely, $[(p-1)!]$, is equal to $[-1]$; therefore, $(p-1)! \equiv -1 \bmod p$.

Conversely, assume that m is composite: there are integers a and b with $m = ab$ and $1 < a \leq b < m$. If $a < b$, then $m = ab$ is a divisor of $(m-1)!$, and so $(m-1)! \equiv 0 \bmod m$. If $a = b$, then $m = a^2$. If $a = 2$, then $(a^2 - 1)! = 3! = 6 \equiv 2 \bmod 4$ and, of course, $2 \not\equiv -1 \bmod 4$. If $2 < a$, then $2a < a^2$, and so a and $2a$ are factors of $(a^2-1)!$; therefore, $(a^2-1)! \equiv 0 \bmod a^2$. Thus, $(a^2-1)! \not\equiv -1 \bmod a^2$, and the proof is complete. •

Remark. We can generalize Wilson's theorem in the same way that Euler's theorem generalizes Fermat's theorem: Replace $U(\mathbb{I}_p)$ by $U(\mathbb{I}_m)$. For example, for all $m \geq 3$, it can be proved that $U(\mathbb{I}_{2^m})$ has exactly 3 elements of order 2, namely, $[-1]$, $[1+2^{m-1}]$, and $[-(1+2^{m-1})]$. It now follows that the product of all the odd numbers r, where $1 \leq r < 2^m$ is congruent to 1 mod 2^m, because

$$\begin{aligned}(-1)(1+2^{m-1})(-1-2^{m-1}) &= (1+2^{m-1})^2 \\ &= 1 + 2^m + 2^{2m-2} \equiv 1 \bmod 2^m. \quad \blacktriangleleft\end{aligned}$$

EXERCISES

2.29 Let H be a subgroup of a group G.
  (i) Prove that right cosets Ha and Hb are equal if and only if $ab^{-1} \in H$.
  (ii) Prove that the relation $a \equiv b$ if $ab^{-1} \in H$ is an equivalence relation on G whose equivalence classes are the right cosets of H.

2.30 (i) Define the ***special linear group*** by

$$\mathrm{SL}(2, \mathbb{R}) = \{A \in \mathrm{GL}(2, \mathbb{R}) : \det(A) = 1\}.$$

  Prove that $\mathrm{SL}(2, \mathbb{R})$ is a subgroup of $\mathrm{GL}(2, \mathbb{R})$.
  (ii) Prove that $\mathrm{GL}(2, \mathbb{Q})$ is a subgroup of $\mathrm{GL}(2, \mathbb{R})$.

2.31 (i) Give an example of two subgroups H and K of a group G whose union $H \cup K$ is not a subgroup of G.
  Hint. Let G be the four-group $\mathbf{V}$.
  (ii) Prove that the union $H \cup K$ of two subgroups is itself a subgroup if and only if either H is a subset of K or K is a subset of H.

2.32 Let G be a finite group with subgroups H and K. If $H \leq K$, prove that

$$[G : H] = [G : K][K : H].$$

2.33 If H and K are subgroups of a group G and if $|H|$ and $|K|$ are relatively prime, prove that $H \cap K = \{1\}$.
Hint. If $x \in H \cap K$, then $x^{|H|} = 1 = x^{|K|}$.

2.34 Prove that every subgroup S of a cyclic group $G = \langle a \rangle$ is itself cyclic.
Hint. If $S \neq 1$, choose k to be the smallest positive integer with $a^k \in S$.

2.35 Prove that a cyclic group G of order n has a subgroup of order d for every d dividing n.
Hint. If $G = \langle a \rangle$ and $n = dk$, consider $\langle a^k \rangle$.

2.36 Let G be a group of order 4. Prove that either G is cyclic or $x^2 = 1$ for every $x \in G$. Conclude, using Exercise 2.26 on page 62, that G must be abelian.

2.37 If H is a subgroup of a group G, prove that the number of left cosets of H in G is equal to the number of right cosets of H in G.
Hint. The function $\varphi : aH \mapsto Ha^{-1}$ is a bijection from the family of all left cosets of H to the family of all right cosets of H.

2.38 Let p be an odd prime, and let $a_1, \ldots, a_{p-1}$ be a permutation of $\{1, 2, \ldots, p-1\}$. Prove that there exist $i \neq j$ with $ia_i \equiv ja_j \bmod p$.
Hint. Use Wilson's theorem.

2.5 HOMOMORPHISMS

An important problem is determining whether two given groups G and H are somehow the same. For example, we have investigated S_3, the group of all permutations of $X = \{1, 2, 3\}$. The group S_Y of all the permutations of $Y = \{a, b, c\}$ is a group different from S_3 because permutations of $\{1, 2, 3\}$ are different than permutations of $\{a, b, c\}$. But even though S_3 and S_Y are different, they surely bear a strong resemblance to each other (see Example 2.51). A more interesting example is the strong resemblance between S_3 and D_6, the symmetries of an equilateral triangle. The notions of homomorphism and isomorphism allow us to compare different groups, as we shall see.

Definition. If $(G, *)$ and $(H, \circ)$ are groups (we have displayed the operation in each), then a function $f : G \to H$ is a ***homomorphism***[11] if

$$f(x * y) = f(x) \circ f(y)$$

for all $x, y \in G$. If f is also a bijection, then f is called an ***isomorphism***. Two groups G and H are called ***isomorphic***, denoted by $G \cong H$, if there exists an isomorphism $f : G \to H$ between them.

A *multiplication table* of a group G displays every product ab for $a, b \in G$.

G	a_1	a_2	$\cdots$	a_j	$\cdots$	a_n
a_1	a_1a_1	a_1a_2	$\cdots$	a_1a_j	$\cdots$	a_1a_n
a_2	a_2a_1	a_2a_2	$\cdots$	a_2a_j	$\cdots$	a_2a_n
a_i	a_ia_1	a_ia_2	$\cdots$	a_ia_j	$\cdots$	a_ia_n
a_n	a_na_1	a_na_2	$\cdots$	a_na_j	$\cdots$	a_na_n

Definition. Let $a_1, a_2, \ldots, a_n$ be a list with no repetitions of all the elements of a group G. A ***multiplication table*** for G is an $n \times n$ array whose ij entry is a_ia_j.

[11]The word *homomorphism* comes from the Greek *homo* meaning "same" and *morph* meaning "shape" or "form." Thus, a homomorphism carries a group to another group (its image) of similar form. The word *isomorphism* involves the Greek *iso* meaning "equal," and isomorphic groups have identical form.

A multiplication table of a group G of order n depends on how we list the elements of G, and so G has $n!$ different multiplication tables. (Thus, the task of determining whether a multiplication table of a group G is the same as some multiplication table of another group H is a daunting one: It involves about $n!$ comparisons, each of which involves comparing n^2 entries.) If $a_1, a_2, \ldots, a_n$ is a list of all the elements of G with no repetitions, and if $f: G \to H$ is a bijection, then $f(a_1), f(a_2), \ldots, f(a_n)$ is a list of all the elements of H with no repetitions, and this latter list determines a multiplication table for H. That f is an isomorphism says that if we superimpose the given multiplication table for G (determined by $a_1, a_2, \ldots, a_n$) upon the multiplication table for H [determined by $f(a_1), f(a_2), \ldots, f(a_n)$], then the tables match: If $a_i a_j$ is the ij entry in the given multiplication table of G, then $f(a_i) f(a_j) = f(a_i a_j)$ is the ij entry of the multiplication table of H. In this sense, isomorphic groups have the same multiplication table. Thus, isomorphic groups are essentially the same, differing only in the notation for the elements and the operations.

Example 2.51.
Let us show that $G = S_3$, the symmetric group permuting $\{1, 2, 3\}$, and $H = S_Y$, the symmetric group of all the permutations of $Y = \{a, b, c\}$, are isomorphic. First, enumerate G:

$$(1), \quad (1\ 2), \quad (1\ 3), \quad (2\ 3), \quad (1\ 2\ 3), \quad (1\ 3\ 2).$$

We define the obvious function $\varphi: S_3 \to S_Y$ that replaces numbers by letters:

$$(1), \quad (a\ b), \quad (a\ c), \quad (b\ c), \quad (a\ b\ c), \quad (a\ c\ b).$$

Compare the multiplication table for S_3 arising from this list of its elements with the multiplication table for S_Y arising from the corresponding list of its elements. The reader should write out the complete tables of each and superimpose one on the other to see that they do match. We will check only one entry. The 4,5 position in the table for S_3 is the product $(2\ 3)(1\ 2\ 3) = (1\ 3)$, while the 4,5 position in the table for S_Y is the product $(b\ c)(a\ b\ c) = (a\ c)$.

This result is generalized in Exercise 2.39 on page 80. ◂

Lemma 2.52. *Let $f: G \to H$ be a homomorphism of groups.*

(i) $f(1) = 1$

(ii) $f(x^{-1}) = f(x)^{-1}$

(iii) $f(x^n) = f(x)^n$ *for all* $n \in \mathbb{Z}$

Sketch of Proof. (i) $1 \cdot 1 = 1$ implies $f(1) f(1) = f(1)$.

(ii) $1 = xx^{-1}$ implies $1 = f(1) = f(x) f(x^{-1})$.

(iii) Use induction to show that $f(x^n) = f(x)^n$ for all $n \geq 0$. Then observe that $x^{-n} = (x^{-1})^n$, and use part (ii). •

Example 2.53.
If G and H are cyclic groups of the same order m, then G and H are isomorphic. (It follows from Corollary 2.45 that any two groups of prime order p are isomorphic.) Although this is not difficult, it requires some care. We have $G = \{1, a, a^2, \ldots, a^{m-1}\}$ and $H = \{1, b, b^2, \ldots, b^{m-1}\}$, and the obvious choice for an isomorphism is the bijection $f : G \to H$ given by $f(a^i) = b^i$. To check that f is an isomorphism, that is, $f(a^i a^j) = b^{i+j}$, involves two cases: $i + j \leq m - 1$; $i + j > m - 1$. We give a less computational proof in Example 2.71. ◀

A property of a group G that is shared by any other group isomorphic to it is called an ***invariant*** of G. For example, the order $|G|$ is an invariant of G, for isomorphic groups have the same orders. Being abelian is an invariant [if f is an isomorphism and a and b commute, then $ab = ba$ and

$$f(a)f(b) = f(ab) = f(ba) = f(b)f(a);$$

hence, $f(a)$ and $f(b)$ commute]. Thus, $\mathbb{I}_6$ and S_3 are not isomorphic, for $\mathbb{I}_6$ is abelian and S_3 is not. In general, it is a challenge to decide whether two given groups are isomorphic. See Exercise 2.42 on page 80 for more examples of invariants.

Example 2.54.
We present two nonisomorphic abelian groups of the same order.

As in Example 2.29(i), let $\mathbf{V}$ be the four-group consisting of the following four permutations:

$$\mathbf{V} = \big\{(1),\ (1\ 2)(3\ 4),\ (1\ 3)(2\ 4),\ (1\ 4)(2\ 3)\big\},$$

and let $\mu_4 = \langle i \rangle = \{1, i, -1, -i\}$ be the multiplicative cyclic group of fourth roots of unity, where $i^2 = -1$. If there were an isomorphism $f : \mathbf{V} \to \mu_4$, then surjectivity of f would provide some $x \in \mathbf{V}$ with $i = f(x)$. But $x^2 = (1)$ for all $x \in \mathbf{V}$, so that $i^2 = f(x)^2 = f(x^2) = f((1)) = 1$, contradicting $i^2 = -1$. Therefore, $\mathbf{V}$ and μ_4 are not isomorphic.

There are other ways to prove this result. For example, μ_4 is cyclic and $\mathbf{V}$ is not; μ_4 has an element of order 4 and $\mathbf{V}$ does not; μ_4 has a unique element of order 2, but $\mathbf{V}$ has 3 elements of order 2. At this stage, you should really believe that μ_4 and $\mathbf{V}$ are not isomorphic! ◀

Definition. If $f : G \to H$ is a homomorphism, define

$$\textbf{\textit{kernel}}^{12}\, f = \{x \in G : f(x) = 1\}$$

and

$$\textbf{\textit{image}}\ f = \{h \in H : h = f(x) \text{ for some } x \in G\}.$$

We usually abbreviate kernel f to ker f and image f to im f.

[12] *Kernel* comes from the German word meaning "grain" or "seed" (*corn* comes from the same word). Its usage here indicates an important ingredient of a homomorphism.

Example 2.55.
(i) If μ_2 is the multiplicative group $\mu_2 = \{\pm 1\}$, then $\operatorname{sgn}\colon S_n \to \mu_2$ is a homomorphism, by Theorem 2.12. The kernel of sgn is the alternating group A_n, the set of all even permutations.

(ii) Determinant is a surjective homomorphism $\det\colon \mathrm{GL}(n, \mathbb{R}) \to \mathbb{R}^\times$, the multiplicative group of nonzero real numbers, whose kernel is the special linear group $\mathrm{SL}(n, \mathbb{R})$ of all $n \times n$ matrices of determinant 1. ◂

Proposition 2.56. *Let $f : G \to H$ be a homomorphism.*

(i) $\ker f$ *is a subgroup of* G *and* $\operatorname{im} f$ *is a subgroup of* H.

(ii) *If* $x \in \ker f$ *and if* $a \in G$, *then* $axa^{-1} \in \ker f$.

(iii) f *is an injection if and only if* $\ker f = \{1\}$.

Sketch of Proof. (i) Routine.

(ii) $f(axa^{-1}) = f(a)1f(a)^{-1} = 1$.

(iii) $f(a) = f(b)$ if and only if $f(b^{-1}a) = 1$. •

Definition. A subgroup K of a group G is called a ***normal subgroup*** if $k \in K$ and $g \in G$ imply $gkg^{-1} \in K$. If K is a normal subgroup of G, we write $K \lhd G$.

The proposition thus says that the kernel of a homomorphism is always a normal subgroup. If G is an abelian group, then every subgroup K is normal, for if $k \in K$ and $g \in G$, then $gkg^{-1} = kgg^{-1} = k \in K$. The converse of this last statement is false: In Example 2.63, we shall show that there is a nonabelian group (the *quaternions*), each of whose subgroups is normal.

The cyclic subgroup $H = \langle(1\ 2)\rangle$ of S_3, consisting of the two elements (1) and (1 2), is not a normal subgroup of S_3: If $\alpha = (1\ 2\ 3)$, then $\alpha^{-1} = (3\ 2\ 1)$, and

$$\alpha(1\ 2)\alpha^{-1} = (1\ 2\ 3)(1\ 2)(3\ 2\ 1) = (2\ 3) \notin H$$

[by Theorem 2.9, $\alpha(1\ 2)\alpha^{-1} = (\alpha 1\ \alpha 2) = (2\ 3)$]. On the other hand, the cyclic subgroup $K = \langle(1\ 2\ 3)\rangle$ of S_3 is a normal subgroup, as the reader should verify.

It follows from Examples 2.55(i) and 2.55(ii) that A_n is a normal subgroup of S_n and $\mathrm{SL}(n, \mathbb{R})$ is a normal subgroup of $\mathrm{GL}(n, \mathbb{R})$ (however, it is also easy to prove these facts directly).

Definition. If G is a group and $a \in G$, then a ***conjugate*** of a is any element in G of the form

$$gag^{-1},$$

where $g \in G$.

It is clear that a subgroup $K \le G$ is a normal subgroup if and only if K contains all the conjugates of its elements: If $k \in K$, then $gkg^{-1} \in K$ for all $g \in G$.

Example 2.57.
(i) Theorem 2.9 states that two permutations in S_n are conjugate if and only if they have the same cycle structure.

(ii) In linear algebra, two matrices $A, B \in \mathrm{GL}(n, \mathbb{R})$ are called ***similar*** if they are conjugate; that is, if there is a nonsingular matrix P with $B = PAP^{-1}$. ◂

Definition. If G is a group and $g \in G$, define ***conjugation*** $\gamma_g : G \to G$ by

$$\gamma_g(a) = gag^{-1}$$

for all $a \in G$.

Proposition 2.58.

(i) *If G is a group and $g \in G$, then conjugation $\gamma_g : G \to G$ is an isomorphism.*

(ii) *Conjugate elements have the same order.*

Proof. (i) If $g, h \in G$, then

$$(\gamma_g \circ \gamma_h)(a) = \gamma_g(hah^{-1}) = g(hah^{-1})g^{-1} = (gh)a(gh)^{-1} = \gamma_{gh}(a);$$

that is,

$$\gamma_g \circ \gamma_h = \gamma_{gh}.$$

It follows that each γ_g is a bijection, for $\gamma_g \circ \gamma_{g^{-1}} = \gamma_1 = 1 = \gamma_{g^{-1}} \circ \gamma_g$. We now show that γ_g is an isomorphism: if $a, b \in G$,

$$\gamma_g(ab) = g(ab)g^{-1} = ga(g^{-1}g)bg^{-1} = \gamma_g(a)\gamma_g(b).$$

(ii) To say that a and b are conjugate is to say that there is $g \in G$ with $b = gag^{-1}$; that is, $b = \gamma_g(a)$. But γ_g is an isomorphism, and so Exercise 2.42 on page 80 shows that a and $b = \gamma_g(a)$ have the same order. •

Example 2.59.
Define the ***center*** of a group G, denoted by $Z(G)$, to be

$$Z(G) = \{z \in G : zg = gz \text{ for all } g \in G\};$$

that is, $Z(G)$ consists of all elements commuting with everything in G.

It is easy to see that $Z(G)$ is a subgroup of G; it is a normal subgroup because if $z \in Z(G)$ and $g \in G$, then

$$gzg^{-1} = zgg^{-1} = z \in Z(G).$$

A group G is abelian if and only if $Z(G) = G$. At the other extreme are ***centerless*** groups G for which $Z(G) = \{1\}$; for example, $Z(S_3) = \{1\}$; indeed, all large symmetric groups are centerless, for Exercise 2.15 on page 51 shows that $Z(S_n) = \{1\}$ for all $n \geq 3$. ◂

Example 2.60.
If G is a group, then an ***automorphism*** of G is an isomorphism $f: G \to G$. For example, every conjugation γ_g is an automorphism of G (it is called an ***inner automorphism***), for its inverse is conjugation by g^{-1}. The set $\operatorname{Aut}(G)$ of all the automorphisms of G is itself a group, under composition, and the set of all conjugations,

$$\operatorname{Inn}(G) = \{\gamma_g : g \in G\},$$

is a subgroup of $\operatorname{Aut}(G)$. Exercise 2.64 on page 82 says that the function $\Gamma: G \to \operatorname{Aut}(G)$, given by $g \mapsto \gamma_g$, is a homomorphism with $\operatorname{im}\Gamma = \operatorname{Inn}(G)$ and $\ker\Gamma = Z(G)$; moreover, $\operatorname{Inn}(G) \lhd \operatorname{Aut}(G)$. ◀

Example 2.61.
The four-group $\mathbf{V}$ is a normal subgroup of S_4. Recall that the elements of $\mathbf{V}$ are

$$\mathbf{V} = \{(1), (1\ 2)(3\ 4), (1\ 3)(2\ 4), (1\ 4)(2\ 3)\}.$$

By Theorem 2.9, every conjugate of a product of two transpositions is another such. But we saw, in Example 2.5(i), that only 3 permutations in S_4 have this cycle structure, and so $\mathbf{V}$ is a normal subgroup of S_4. ◀

Proposition 2.62.

(i) *If H is a subgroup of index 2 in a group G, then $g^2 \in H$ for every $g \in G$.*

(ii) *If H is a subgroup of index 2 in a group G, then H is a normal subgroup of G.*

Proof. (i) Since H has index 2, there are exactly two cosets, namely, H and aH, where $a \notin H$. Thus, G is the disjoint union $G = H \,\dot\cup\, aH$. Take $g \in G$ with $g \notin H$, so that $g = ah$ for some $h \in H$. If $g^2 \notin H$, then $g^2 = ah'$, where $h' \in H$. Hence,

$$g = g^{-1}g^2 = h^{-1}a^{-1}ah' = h^{-1}h' \in H,$$

and this is a contradiction.

(ii) [13] It suffices to prove that if $h \in H$, then the conjugate $ghg^{-1} \in H$ for every $g \in G$. Since H has index 2, there are exactly two cosets, namely, H and aH, where $a \notin H$. Now, either $g \in H$ or $g \in aH$. If $g \in H$, then $ghg^{-1} \in H$, because H is a subgroup. In the second case, write $g = ax$, where $x \in H$. Then $ghg^{-1} = a(xhx^{-1})a^{-1} = ah'a^{-1}$, where $h' = xhx^{-1} \in H$ (for h' is a product of three elements in H). If $ghg^{-1} \notin H$, then $ghg^{-1} = ah'a^{-1} \in aH$; that is, $ah'a^{-1} = ay$ for some $y \in H$. Canceling a, we have $h'a^{-1} = y$, which gives the contradiction $a = y^{-1}h' \in H$. Therefore, if $h \in H$, every conjugate of h also lies in H; that is, H is a normal subgroup of G. •

[13] Another proof of this is given in Exercise 2.50 on page 81.

Definition. The group of ***quaternions***[14] is the group **Q** of order 8 consisting of the following matrices in $\mathrm{GL}(2, \mathbb{C})$:

$$\mathbf{Q} = \{\, I, A, A^2, A^3, B, BA, BA^2, BA^3 \,\},$$

where I is the identity matrix,

$$A = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}, \text{ and } B = \begin{bmatrix} 0 & i \\ i & 0 \end{bmatrix}.$$

The element $A \in \mathbf{Q}$ has order 4, so that $\langle A \rangle$ is a subgroup of order 4 and hence of index 2; the other coset is $B\langle A \rangle = \{B, BA, BA^2, BA^3\}$. Thus, every element in **Q** has an expression of the form $B^i A^j$, where $i = 0, 1$ and $j = 0, 1, 2, 3$.

Example 2.63.
In Exercise 2.59 on page 81, the reader will check that **Q** is a nonabelian group of order 8 having exactly one element of order 2, and hence only one subgroup of order 2, namely, $\langle -I \rangle$. We claim that every subgroup of **Q** is normal. Lagrange's theorem says that every subgroup of Q has order a divisor of 8, and so the only possible orders of subgroups are 1, 2, 4, or 8. Clearly, the subgroup $\{I\}$ and the subgroup of order 8 (namely, **Q** itself) are normal subgroups. By Proposition 2.62(ii), any subgroup of order 4 must be normal, for it has index 2. Finally, the subgroup $\langle -I \rangle$ is normal, for it is the center, $Z(\mathbf{Q})$. ◄

Example 2.63 shows that **Q** is a nonabelian group that is like abelian groups in that every subgroup is normal. This is essentially the only such example. A nonabelian finite group is called ***hamiltonian*** if every subgroup is normal; every hamiltonian group has the form $\mathbf{Q} \times A$, where A is an abelian group with no elements of order 4 (*direct products* will be introduced in the next section). A proof of this result can be found in Robinson, *A Course in the Theory of Groups*, page 139.

Lagrange's theorem states that the order of a subgroup of a finite group G must be a divisor of $|G|$. This suggests the question, given a divisor d of $|G|$, whether G must contain a subgroup of order d. The next result shows that there need not be such a subgroup.

Proposition 2.64. *The alternating group A_4 is a group of order* 12 *having no subgroup of order* 6*.*

Proof. First, $|A_4| = 12$, by Exercise 2.12 on page 50. If A_4 contains a subgroup H of order 6, then H has index 2, and so $\alpha^2 \in H$ for every $\alpha \in A_4$, by Corollary 2.62(i). If α is a 3-cycle, however, then α has order 3, so that $\alpha = \alpha^4 = (\alpha^2)^2$. Thus, H contains every 3-cycle. This is a contradiction, for there are 8 3-cycles in A_4. •

[14] W. R. Hamilton invented a system having two operations, addition and multiplication, that he called quaternions, for it was four-dimensional. The group of quaternions consists of 8 special elements in that system; see Exercise 2.60 on page 82.

EXERCISES

2.39 Show that if there is a bijection $f: X \to Y$ (that is, if X and Y have the same number of elements), then there is an isomorphism $\varphi: S_X \to S_Y$.
Hint. If $\alpha \in S_X$, define $\varphi(\alpha) = f \circ \alpha \circ f^{-1}$. In particular, show that if $|X| = 3$, then φ takes a cycle involving symbols 1, 2, 3 into a cycle involving a, b, c, as in Example 2.51.

2.40 (i) Show that the composite of homomorphisms is itself a homomorphism.
(ii) Show that the inverse of an isomorphism is an isomorphism.
(iii) Show that two groups that are isomorphic to a third group are isomorphic to each other.
(iv) Prove that isomorphism is an equivalence relation on any set of groups.

2.41 Prove that a group G is abelian if and only if the function $f: G \to G$, given by $f(a) = a^{-1}$, is a homomorphism.

2.42 This exercise gives some invariants of a group G. Let $f: G \to H$ be an isomorphism.
(i) Prove that if $a \in G$ has infinite order, then so does $f(a)$, and if a has finite order n, then so does $f(a)$. Conclude that if G has an element of some order n and H does not, then $G \ncong H$.
(ii) Prove that if $G \cong H$, then, for every divisor d of $|G|$, both G and H have the same number of elements of order d.

2.43 Prove that A_4 and D_{12} are nonisomorphic groups of order 12.

2.44 (i) Find a subgroup H of S_4 with $H \neq \mathbf{V}$ and $H \cong \mathbf{V}$.
(ii) Prove that the subgroup H in part (i) is not a normal subgroup.

2.45 Show that every group G with $|G| < 6$ is abelian.

2.46 Let $G = \{f: \mathbb{R} \to \mathbb{R} : f(x) = ax + b, \text{ where } a \neq 0\}$. Prove that G is a group under composition that is isomorphic to the subgroup of $\mathrm{GL}(2, \mathbb{R})$ consisting of all matrices of the form $\begin{bmatrix} a & b \\ 0 & 1 \end{bmatrix}$.

2.47 (i) If $f: G \to H$ is a homomorphism and $x \in G$ has order k, prove that $f(x) \in H$ has order m, where $m \mid k$.
(ii) If $f: G \to H$ is a homomorphism and if $(|G|, |H|) = 1$, prove that $f(x) = 1$ for all $x \in G$.

2.48 (i) Prove that

$$\begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}^k = \begin{bmatrix} \cos k\theta & -\sin k\theta \\ \sin k\theta & \cos k\theta \end{bmatrix}.$$

Hint. Use induction on $k \geq 1$.
(ii) Prove that the special orthogonal group $SO(2, \mathbb{R})$, consisting of all 2×2 orthogonal matrices of determinant 1, is isomorphic to the circle group S^1.
Hint. Consider $\varphi : \begin{bmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{bmatrix} \mapsto (\cos\alpha, \sin\alpha)$.

2.49 Let G be the additive group of all polynomials in x with coefficients in $\mathbb{Z}$, and let H be the multiplicative group of all positive rationals. Prove that $G \cong H$.
Hint. List the prime numbers $p_0 = 2$, $p_1 = 3$, $p_2 = 5, \ldots,$ and define

$$\varphi(e_0 + e_1 x + e_2 x^2 + \cdots + e_n x^n) = p_0^{e_0} \cdots p_n^{e_n}.$$

2.50 (i) Show that if H is a subgroup with $bH = Hb = \{hb : h \in H\}$ for every $b \in G$, then H must be a normal subgroup.

(ii) Use part (i) to give a second proof of Proposition 2.62(ii): If $H \leq G$ has index 2, then $H \lhd G$.

Hint. If $a \notin H$, then $aH = H' = Ha$, where H' is the complement of H.

2.51 (i) Prove that if $\alpha \in S_n$, then α and α^{-1} are conjugate.

(ii) Give an example of a group G containing an element x for which x and x^{-1} are not conjugate.

2.52 Prove that the intersection of any family of normal subgroups of a group G is itself a normal subgroup of G.

2.53 Define $W = \langle(1\ 2)(3\ 4)\rangle$, the cyclic subgroup of S_4 generated by $(1\ 2)(3\ 4)$. Show that W is a normal subgroup of $\mathbf{V}$, but that W is not a normal subgroup of S_4. Conclude that normality is not transitive: $W \lhd \mathbf{V}$ and $\mathbf{V} \lhd G$ do not imply $W \lhd G$.

2.54 Let G be a finite abelian group written multiplicatively. Prove that if $|G|$ is odd, then every $x \in G$ has a unique square root; that is, there exists exactly one $g \in G$ with $g^2 = x$.
Hint. Show that squaring is an injective function $G \to G$, and use Exercise 1.58 on page 36.

2.55 Give an example of a group G, a subgroup $H \leq G$, and an element $g \in G$ with $[G : H] = 3$ and $g^3 \notin H$.
Hint. Take $G = S_3$, $H = \langle(1\ 2)\rangle$, and $g = (2\ 3)$.

2.56 Show that the center of $\mathrm{GL}(2, \mathbb{R})$ is the set of all ***scalar matrices*** aI with $a \neq 0$.
Hint. Show that if A is a matrix that is not a scalar matrix, then there is some nonsingular matrix that does not commute with A. (The generalization of this to $n \times n$ matrices is true.)

2.57 Let $\zeta = e^{2\pi i/n}$ be a primitive nth root of unity, and define

$$A = \begin{bmatrix} \zeta & 0 \\ 0 & \zeta^{-1} \end{bmatrix} \text{ and } B = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}.$$

(i) Prove that A has order n and that B has order 2.

(ii) Prove that $BAB = A^{-1}$.

(iii) Prove that the matrices of the form A^i and BA^i, for $0 \leq i < n$, form a multiplicative subgroup $G \leq \mathrm{GL}(2, \mathbb{C})$.

Hint. Consider cases A^iA^j, A^iBA^j, BA^iA^j, and $(BA^i)(BA^j)$.

(iv) Prove that each matrix in G has a unique expression of the form B^iA^j, where $i = 0, 1$ and $0 \leq j < n$. Conclude that $|G| = 2n$.

(v) Prove that $G \cong D_{2n}$.

Hint. Define a function $G \to D_{2n}$ using the unique expression of elements in G in the form B^iA^j.

2.58 (i) Prove that every subgroup of $\mathbf{Q} \times \mathbb{I}_2$ is normal.

(ii) Prove that there exists a nonnormal subgroup of $\mathbf{Q} \times \mathbb{I}_4$.

2.59 Recall that the group of quaternions $\mathbf{Q}$ consists of the 8 matrices in $\mathrm{GL}(2, \mathbb{C})$

$$\mathbf{Q} = \{\, I, A, A^2, A^3, B, BA, BA^2, BA^3 \,\},$$

where

$$A = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \text{ and } B = \begin{bmatrix} 0 & i \\ i & 0 \end{bmatrix}.$$

(i) Prove that $-I$ is the only element in $\mathbf{Q}$ of order 2, and that all other elements $M \neq I$ satisfy $M^2 = -I$.

(ii) Prove that $\mathbf{Q}$ is a nonabelian group with operation matrix multiplication.
Hint. Note that $A^2 = -I = B^2$.

(iii) Prove that $\mathbf{Q}$ has a unique subgroup of order 2, and it is the center of $\mathbf{Q}$.

2.60 Assume that there is a group G of order 8 whose elements

$$\pm 1, \pm \mathbf{i}, \pm \mathbf{j}, \pm \mathbf{k}$$

satisfy

$$\mathbf{i}^2 = \mathbf{j}^2 = \mathbf{k}^2 = -1, \qquad \mathbf{ij} = \mathbf{k}, \qquad \mathbf{jk} = \mathbf{i}, \qquad \mathbf{ki} = \mathbf{j},$$
$$\mathbf{ij} = -\mathbf{ji}, \qquad \mathbf{ik} = -\mathbf{ki}, \qquad \mathbf{jk} = -\mathbf{kj}.$$

Prove that $G \cong \mathbf{Q}$ and, conversely, that $\mathbf{Q}$ is such a group.

2.61 Prove that the quaternions $\mathbf{Q}$ and the dihedral group D_8 are nonisomorphic groups of order 8.
Hint. Use Exercise 2.42 on page 80.

2.62 Prove that A_4 is the only subgroup of S_4 of order 12.
Hint. Use Proposition 2.62(ii).

2.63 Prove that the symmetry group $\Sigma(\pi_n)$, where π_n is a regular polygon with n vertices, is isomorphic to a subgroup of S_n.
Hint. The vertices $X = \{v_1, \ldots, v_n\}$ of π_n are permuted by every motion $\sigma \in \Sigma(\pi_n)$.

2.64 (i) For every group G, show that the function $\Gamma \colon G \to \mathrm{Aut}(G)$, given by $g \mapsto \gamma_g$ (where γ_x is conjugation by g), is a homomorphism.

(ii) Prove that $\ker \Gamma = Z(G)$ and $\operatorname{im} \Gamma = \mathrm{Inn}(G)$; conclude that $\mathrm{Inn}(G)$ is a subgroup of $\mathrm{Aut}(G)$.

(iii) Prove that $\mathrm{Inn}(G) \lhd \mathrm{Aut}(G)$.

2.6 Quotient Groups

The construction of the additive group of integers modulo m is the prototype of a more general way of building new groups from given groups, called *quotient groups*. The homomorphism $\pi \colon \mathbb{Z} \to \mathbb{I}_m$, defined by $\pi \colon a \mapsto [a]$, is surjective, so that $\mathbb{I}_m$ is equal to $\operatorname{im} \pi$. Thus, every element of $\mathbb{I}_m$ has the form $\pi(a)$ for some $a \in \mathbb{Z}$, and $\pi(a) + \pi(b) = \pi(a+b)$. This description of the additive group $\mathbb{I}_m$ in terms of the additive group $\mathbb{Z}$ can be generalized to arbitrary, not necessarily abelian, groups. Suppose that $f \colon G \to H$ is a surjective homomorphism between groups G and H. Since f is surjective, each element of H has the form $f(a)$ for some $a \in G$, and the operation in H is given by $f(a)f(b) = f(ab)$, where

$a, b \in G$. Now $K = \ker f$ is a normal subgroup of G, and we are going to reconstruct $H = \operatorname{im} f$ (as well as a surjective homomorphism $\pi : G \to H$) from G and K alone.

We begin by introducing an operation on the set

$$\mathcal{S}(G)$$

of all nonempty subsets of a group G. If $X, Y \in \mathcal{S}(G)$, define

$$XY = \{xy : x \in X \text{ and } y \in Y\}.$$

This multiplication is associative: $X(YZ)$ is the set of all $x(yz)$, where $x \in X$, $y \in Y$, and $z \in Z$, $(XY)Z$ is the set of all such $(xy)z$, and these are the same because of associativity in G.

An instance of this multiplication is the product of a one-point subset $\{a\}$ and a subgroup $K \leq G$, which is the coset aK.

As a second example, we show that if H is any subgroup of G, then

$$HH = H.$$

If $h, h' \in H$, then $hh' \in H$, because subgroups are closed under multiplication, and so $HH \subseteq H$. For the reverse inclusion, if $h \in H$, then $h = h1 \in HH$ (because $1 \in H$), and so $H \subseteq HH$.

It is possible for two subsets X and Y in $\mathcal{S}(G)$ to commute even though their constituent elements do not commute. For example, let $G = S_3$ and $K = \langle(1\ 2\ 3)\rangle$. Now $(1\ 2)$ does not commute with $(1\ 2\ 3) \in K$, but we claim that $(1\ 2)K = K(1\ 2)$. In fact, here is the converse of Exercise 2.50 on page 81.

Lemma 2.65. *A subgroup K of a group G is a normal subgroup if and only if*

$$gK = Kg$$

for every $g \in G$. Thus, every right coset of a normal subgroup is also a left coset.

Proof. Let $gk \in gK$. Since K is normal, $gkg^{-1} \in K$, say $gkg^{-1} = k' \in K$, so that $gk = (gkg^{-1})g = k'g \in Kg$, and so $gK \subseteq Kg$. For the reverse inclusion, let $kg \in Kg$. Since K is normal, $(g^{-1})k(g^{-1})^{-1} = g^{-1}kg \in K$, say $g^{-1}kg = k'' \in K$. Hence, $kg = g(g^{-1}kg) = gk'' \in gK$ and $Kg \subseteq gK$. Therefore, $gK = Kg$ when $K \lhd G$.

Conversely, if $gK = Kg$ for every $g \in G$, then for each $k \in K$, there is $k' \in K$ with $gk = k'g$; that is, $gkg^{-1} \in K$ for all $g \in G$, and so $K \lhd G$. •

A natural question is whether HK is a subgroup when both H and K are subgroups. In general, HK need not be a subgroup. For example, let $G = S_3$, let $H = \langle(1\ 2)\rangle$, and let $K = \langle(1\ 3)\rangle$. Then

$$HK = \{(1), (1\ 2), (1\ 3), (1\ 3\ 2)\}$$

is not a subgroup lest we contradict Lagrange's theorem, for $4 \nmid 6$.

Proposition 2.66.

(i) *If H and K are subgroups of a group G, and if one of them is a normal subgroup, then HK is a subgroup of G; moreover, $HK = KH$ in this case.*

(ii) *If both H and K are normal subgroups, then HK is a normal subgroup.*

Remark. Exercise 2.72 on page 95 shows that if H and K are subgroups of a group G, then HK is a subgroup if and only if $HK = KH$. ◂

Proof. (i) Assume first that $K \lhd G$. We claim that $HK = KH$. If $hk \in HK$, then $k' = hkh^{-1} \in K$, because $K \lhd G$, and

$$hk = hkh^{-1}h = k'h \in KH.$$

Hence, $HK \subseteq KH$. For the reverse inclusion, write $kh = hh^{-1}kh = hk'' \in HK$. (Note that the same argument shows that $HK = KH$ if $H \lhd G$.)

We now show that HK is a subgroup. Since $1 \in H$ and $1 \in K$, we have $1 = 1 \cdot 1 \in HK$; if $hk \in HK$, then $(hk)^{-1} = k^{-1}h^{-1} \in KH = HK$; if $hk, h_1k_1 \in HK$, then $hkh_1k_1 \in HKHK = HHKK = HK$.

(ii) If $g \in G$, then Lemma 2.65 gives $gHK = HgK = HKg$, and the same lemma now gives $HK \lhd G$. •

Here is a fundamental construction of a new group from a given group.

Theorem 2.67. *Let G/K denote the family of all the left cosets of a subgroup K of G. If K is a normal subgroup, then*

$$aKbK = abK$$

for all $a, b \in G$, and G/K is a group under this operation.

Remark. The group G/K is called the ***quotient group*** G mod K; when G is finite, its order $|G/K|$ is the index $[G : K] = |G|/|K|$ (presumably, this is the reason why *quotient groups* are so called). ◂

Proof. The product of two cosets $(aK)(bK)$ can also be viewed as the product of 4 elements in $\mathcal{S}(G)$. Hence, associativity in $\mathcal{S}(G)$ gives

$$(aK)(bK) = a(Kb)K = a(bK)K = abKK = abK,$$

for normality of K gives $Kb = bK$ for all $b \in K$, by Lemma 2.65, while $KK = K$ because K is a subgroup. Thus, the product of two cosets of K is again a coset of K, and so an operation on G/K has been defined. Because multiplication in $\mathcal{S}(G)$ is associative, equality $X(YZ) = (XY)Z$ holds, in particular, when X, Y, and Z are cosets of K, so that the operation on G/K is associative. The identity is the coset $K = 1K$, for $(1K)(bK) = 1bK = bK = b1K = (bK)(1K)$, and the inverse of aK is $a^{-1}K$, for $(a^{-1}K)(aK) = a^{-1}aK = K = aa^{-1}K = (aK)(a^{-1}K)$. Therefore, G/K is a group. •

It is important to remember what we have just proved: The product $aKbK = abK$ in G/K does not depend on the particular representatives of the cosets, and the law of substitution holds: If $aK = a'K$ and $bK = b'K$, then

$$aKbK = abK = a'b'K = a'Kb'K.$$

Example 2.68.
We show that the quotient group G/K is precisely $\mathbb{I}_m$ when G is the additive group $\mathbb{Z}$ and $K = \langle m \rangle$, the (cyclic) subgroup of all the multiples of a positive integer m. Since $\mathbb{Z}$ is abelian, $\langle m \rangle$ is necessarily a normal subgroup. The sets $\mathbb{Z}/\langle m \rangle$ and $\mathbb{I}_m$ coincide because they are comprised of the same elements: The coset $a + \langle m \rangle$ is the congruence class $[a]$:

$$a + \langle m \rangle = \{a + km : k \in \mathbb{Z}\} = [a].$$

The operations also coincide: Addition in $\mathbb{Z}/\langle m \rangle$ is given by

$$(a + \langle m \rangle) + (b + \langle m \rangle) = (a + b) + \langle m \rangle;$$

since $a + \langle m \rangle = [a]$, this last equation is just $[a] + [b] = [a + b]$, which is the sum in $\mathbb{I}_m$. Therefore, $\mathbb{I}_m$ is equal to the quotient group $\mathbb{Z}/\langle m \rangle$. ◂

There is another way to regard quotient groups. After all, we saw, in the proof of Lemma 2.40, that the relation $\equiv$ on G, defined by $a \equiv b$ if $b^{-1}a \in K$, is an equivalence relation whose equivalence classes are the cosets of K. Thus, we can view the elements of G/K as equivalence classes, with the multiplication $aKbK = abK$ being independent of the choice of representative.

We remind the reader of Lemma 2.40(i): If K is a subgroup of G, then two cosets aK and bK are equal if and only if $b^{-1}a \in K$. In particular, if $b = 1$, then $aK = K$ if and only if $a \in K$.

We can now prove the converse of Proposition 2.56(ii).

Corollary 2.69. *Every normal subgroup $K \lhd G$ is the kernel of some homomorphism.*

Proof. Define the ***natural map*** $\pi\colon G \to G/K$ by $\pi(a) = aK$. With this notation, the formula $aKbK = abK$ can be rewritten as $\pi(a)\pi(b) = \pi(ab)$; thus, π is a (surjective) homomorphism. Since K is the identity element in G/K,

$$\ker \pi = \{a \in G : \pi(a) = K\} = \{a \in G : aK = K\} = K,$$

by Lemma 2.40(i). •

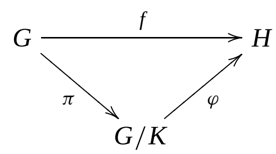
The next theorem shows that every homomorphism gives rise to an isomorphism and that quotient groups are merely constructions of homomorphic images. E. Noether (1882–1935) emphasized the fundamental importance of this fact.

Theorem 2.70 (First Isomorphism Theorem). *If $f\colon G \to H$ is a homomorphism, then*

$$\ker f \lhd G \quad \textit{and} \quad G/\ker f \cong \operatorname{im} f.$$

In more detail, if $\ker f = K$ *and* $\varphi\colon G/K \to \operatorname{im} f \le H$ *is given by* $\varphi\colon aK \mapsto f(a)$, *then* φ *is an isomorphism.*

Remark. The following diagram describes the proof of the first isomorphism theorem, where $\pi : G \to G/K$ is the natural map $\pi : a \mapsto aK$.



◄

Proof. We have already seen, in Proposition 2.56(ii), that $K = \ker f$ is a normal subgroup of G. Now φ is well-defined: If $aK = bK$, then $a = bk$ for some $k \in K$, and so $f(a) = f(bk) = f(b)f(k) = f(b)$, because $f(k) = 1$.

Let us now see that φ is a homomorphism. Since f is a homomorphism and $\varphi(aK) = f(a)$,

$$\varphi(aKbK) = \varphi(abK) = f(ab) = f(a)f(b) = \varphi(aK)\varphi(bK).$$

It is clear that $\operatorname{im}\varphi \le \operatorname{im} f$. For the reverse inclusion, note that if $y \in \operatorname{im} f$, then $y = f(a)$ for some $a \in G$, and so $y = f(a) = \varphi(aK)$. Thus, φ is surjective.

Finally, we show that φ is injective. If $\varphi(aK) = \varphi(bK)$, then $f(a) = f(b)$. Hence, $1 = f(b)^{-1}f(a) = f(b^{-1}a)$, so that $b^{-1}a \in \ker f = K$. Therefore, $aK = bK$, by Lemma 2.40(i), and so φ is injective. We have proved that $\varphi\colon G/K \to \operatorname{im} f$ is an isomorphism. •

Given any homomorphism $f : G \to H$, we should immediately ask for its kernel and image; the first isomorphism theorem will then provide an isomorphism $G/\ker f \cong \operatorname{im} f$. Since there is no significant difference between isomorphic groups, the first isomorphism theorem also says that there is no significant difference between quotient groups and homomorphic images.

Example 2.71.
Let us revisit Example 2.53, which showed that any two cyclic groups of order m are isomorphic. Let $G = \langle a \rangle$ be a cyclic group of order m. Define a function $f : \mathbb{Z} \to G$ by $f(n) = a^n$ for all $n \in \mathbb{Z}$. Now f is easily seen to be a homomorphism; it is surjective (because a is a generator of G), while $\ker f = \{n \in \mathbb{Z} : a^n = 1\} = \langle m \rangle$, by Theorem 2.24. The first isomorphism theorem gives an isomorphism $\mathbb{Z}/\langle m \rangle \cong G$. We have shown that every cyclic group of order m is isomorphic to $\mathbb{Z}/\langle m \rangle$, and hence that any two cyclic groups of order m are isomorphic to each other. Of course, Example 2.68 shows that $\mathbb{Z}/\langle m \rangle = \mathbb{I}_m$, so that every cyclic group of order m is isomorphic to $\mathbb{I}_m$.

We point out that any two infinite cyclic groups are isomorphic to $\mathbb{Z}$; the reader should have no difficulty proving this. ◄

Example 2.72.
What is the quotient group $\mathbb{R}/\mathbb{Z}$? Define $f : \mathbb{R} \to S^1$, where S^1 is the circle group, by

$$f : x \mapsto e^{2\pi i x}.$$

Now f is a homomorphism; that is, $f(x+y) = f(x)f(y)$, by the addition formulas for sine and cosine. The map f is surjective, and $\ker f$ consists of all $x \in \mathbb{R}$ for which $e^{2\pi ix} = \cos 2\pi x + i \sin 2\pi x = 1$; that is, $\cos 2\pi x = 1$ and $\sin 2\pi x = 0$. But $\cos 2\pi x = 1$ forces x to be an integer; since $1 \in \ker f$, we have $\ker f = \mathbb{Z}$. The first isomorphism theorem now gives

$$\mathbb{R}/\mathbb{Z} \cong S^1.$$

This is the group-theoretic version of Example 1.55(i). ◂

Here is a useful counting result.

Proposition 2.73 (Product Formula). *If H and K are subgroups of a finite group G, then*

$$|HK||H \cap K| = |H||K|,$$

where $HK = \{hk : h \in H$ and $k \in K\}$.

Remark. The subset HK need not be a subgroup of G; however, Proposition 2.66 shows that if either $H \lhd G$ or $K \lhd G$, then HK is a subgroup (see also Exercise 2.72 on page 95). ◂

Proof. Define a function $f\colon H \times K \to HK$ by $f : (h,k) \mapsto hk$. Clearly, f is a surjection. It suffices to show, for every $x \in HK$, that $|f^{-1}(x)| = |H \cap K|$, where $f^{-1}(x) = \{(h,k) \in H \times K : hk = x\}$, [because $H \times K$ is the disjoint union $\bigcup_{x \in HK} f^{-1}(x)$].

We claim that if $x = hk$, then

$$f^{-1}(x) = \{(hd, d^{-1}k) : d \in H \cap K\}.$$

Each $(hd, d^{-1}k) \in f^{-1}(x)$, for $f(hd, d^{-1}k) = hdd^{-1}k = hk = x$. For the reverse inclusion, let $(h', k') \in f^{-1}(x)$, so that $h'k' = hk$. Then $h^{-1}h' = kk'^{-1} \in H \cap K$; call this element d. Then $h' = hd$ and $k' = d^{-1}k$, and so (h', k') lies in the right side. Therefore,

$$|f^{-1}(x)| = |\{(hd, d^{-1}k) : d \in H \cap K\}| = |H \cap K|,$$

because $d \mapsto (hd, d^{-1}k)$ is a bijection. •

The next two results are consequences of the first isomorphism theorem.

Theorem 2.74 (Second Isomorphism Theorem). *If H and K are subgroups of a group G with $H \lhd G$, then HK is a subgroup, $H \cap K \lhd K$, and*

$$K/(H \cap K) \cong HK/H.$$

Proof. Since $H \lhd G$, Proposition 2.66 shows that HK is a subgroup. Normality of H in HK follows from a more general fact: If $H \le S \le G$ and if H is normal in G, then H is normal in S (if $ghg^{-1} \in H$ for every $g \in G$, then, in particular, $ghg^{-1} \in H$ for every $g \in S$).

We now show that every coset $xH \in HK/H$ has the form kH for some $k \in K$. Of course, $xH = hkH$, where $h \in H$ and $k \in K$. But $hk = kk^{-1}hk = kh'$ for some $h' \in H$, so that $hkH = kh'H = kH$. It follows that the function $f : K \to HK/H$, given by $f : k \mapsto kH$, is surjective. Moreover, f is a homomorphism, for it is the restriction of the natural map $\pi : G \to G/H$. Since $\ker \pi = H$, it follows that $\ker f = H \cap K$, and so $H \cap K$ is a normal subgroup of K. The first isomorphism theorem now gives $K/(H \cap K) \cong HK/H$. $\bullet$

The second isomorphism theorem gives the product formula in the special case when one of the subgroups is normal: If $K/(H \cap K) \cong HK/H$, then $|K/(H \cap K)| = |HK/H|$, and so $|HK||H \cap K| = |H||K|$.

Theorem 2.75 (Third Isomorphism Theorem). *If H and K are normal subgroups of a group G with $K \le H$, then $H/K \lhd G/K$ and*

$$(G/K)/(H/K) \cong G/H.$$

Proof. Define $f : G/K \to G/H$ by $f : aK \mapsto aH$. Note that f is a (well-defined) function, for if $a' \in G$ and $a'K = aK$, then $a^{-1}a' \in K \le H$, and so $aH = a'H$. It is easy to see that f is a surjective homomorphism.

Now $\ker f = H/K$, for $aH = H$ if and only if $a \in H$, and so H/K is a normal subgroup of G/K. Since f is surjective, the first isomorphism theorem gives

$$(G/K)/(H/K) \cong G/H. \quad \bullet$$

The third isomorphism theorem is easy to remember: In the fraction $(G/K)/(H/K)$, the K's can be canceled. We can better appreciate the first isomorphism theorem after having proved the third one. The quotient group $(G/K)/(H/K)$ consists of cosets (of H/K) whose representatives are themselves cosets (of G/K). A direct proof of the third isomorphism theorem could be nasty.

The next result, which can be regarded as a fourth isomorphism theorem, describes the subgroups of a quotient group G/K.

Proposition 2.76 (Correspondence Theorem). *Let G be a group, let $K \lhd G$, and let $\pi : G \to G/K$ be the natural map. Then*

$$S \mapsto \pi(S) = S/K$$

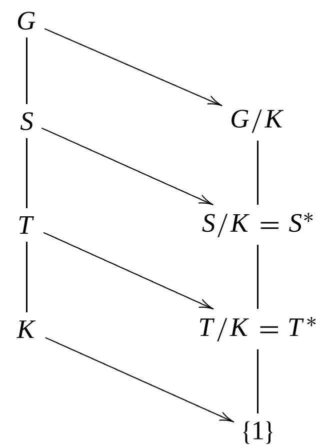
is a bijection between $\operatorname{Sub}(G; K)$*, the family of all those subgroups S of G that contain K, and* $\operatorname{Sub}(G/K)$*, the family of all the subgroups of G/K. If we denote S/K by S^*, then*

$$T \le S \le G \quad \textit{if and only if} \quad T^* \le S^*, \ \textit{in which case } [S : T] = [S^* : T^*],$$

and

$$T \lhd S \quad \textit{if and only if} \quad T^* \lhd S^*, \ \textit{in which case } S/T \cong S^*/T^*.$$

Remark. The following diagram is a way to remember this theorem.



◂

Proof. Define $\Phi\colon \mathrm{Sub}(G; K) \to \mathrm{Sub}(G/K)$ by $\Phi\colon S \mapsto S/K$ (it is routine to check that if S is subgroup of G containing K, then S/K is a subgroup of G/K).

To see that Φ is injective, we begin by showing that if $K \le S \le G$, then $\pi^{-1}\pi(S) = S$. As always, $S \subseteq \pi^{-1}\pi(S)$, by Proposition 1.50(iv). For the reverse inclusion, let $a \in \pi^{-1}\pi(S)$, so that $\pi(a) = \pi(s)$ for some $s \in S$. It follows that $as^{-1} \in \ker\pi = K$, so that $a = sk$ for some $k \in K$. But $K \le S$, and so $a = sk \in S$.

Assume now that $\pi(S) = \pi(S')$, where S and S' are subgroups of G containing K. Then $\pi^{-1}\pi(S) = \pi^{-1}\pi(S')$, and so $S = S'$ as we have just proved in the preceding paragraph; hence, Φ is injective.

To see that Φ is surjective, let U be a subgroup of G/K. Now $\pi^{-1}(U)$ is a subgroup of G containing $K = \pi^{-1}(\{1\})$, and $\pi(\pi^{-1}(U)) = U$, by Proposition 1.50(ii).

Proposition 1.50(i) shows that $T \le S \le G$ implies $T/K = \pi(T) \le \pi(S) = S/K$. Conversely, assume that $T/K \le S/K$. If $t \in T$, then $tK \in T/K \le S/K$ and so $tK = sK$ for some $s \in S$. Hence, $t = sk$ for some $k \in K \le S$, and so $t \in S$.

To prove that $[S : T] = [S^* : T^*]$, it suffices to show that there is a bijection from the family of all cosets of the form sT, where $s \in S$, and the family of all cosets of the form s^*T^*, where $s^* \in S^*$, and the reader may check that $sT \mapsto \pi(s)T^*$ is such a bijection. When G is finite, we may prove $[S : T] = [S^* : T^*]$ as follows:

$$\begin{aligned}
[S^* : T^*] &= |S^*|/|T^*| \\
&= |S/K|/|T/K| \\
&= (|S|/|K|)\,/\,(|T|/|K|) \\
&= |S|/|T| \\
&= [S : T].
\end{aligned}$$

If $T \lhd S$, then $T/K \lhd S/K$ and $(S/K)/(T/K) \cong S/T$, by the third isomorphism theorem; that is, $S^*/T^* \cong S/T$. It remains to show that if $T^* \lhd S^*$, then $T \lhd S$; that is, if $t \in T$ and $s \in S$, then $sts^{-1} \in T$. Now

$$\pi(sts^{-1}) = \pi(s)\pi(t)\pi(s)^{-1} \in \pi(s)T^*\pi(s)^{-1} = T^*,$$

so that $sts^{-1} \in \pi^{-1}(T^*) = T$. •

When dealing with quotient groups, we usually say, without mentioning the correspondence theorem explicitly, that every subgroup of G/K has the form S/K for a unique subgroup $S \le G$ containing K.

Example 2.77.
Let $G = \langle a \rangle$ be a cyclic group of order 30. If $\pi : \mathbb{Z} \to G$ is defined by $\pi(n) = a^n$, then $\ker \pi = \langle 30 \rangle$. The subgroups $\langle 30 \rangle \le \langle 15 \rangle \le \langle 5 \rangle \le \mathbb{Z}$ correspond to the subgroups

$$\{1\} = \langle a^{30} \rangle \le \langle a^{15} \rangle \le \langle a^5 \rangle \le \langle a \rangle.$$

Moreover, the quotient groups are

$$\frac{\langle a^{15} \rangle}{\langle a^{30} \rangle} \cong \frac{\langle 15 \rangle}{\langle 30 \rangle} \cong \mathbb{I}_2, \quad \frac{\langle a^{5} \rangle}{\langle a^{15} \rangle} \cong \frac{\langle 5 \rangle}{\langle 15 \rangle} \cong \mathbb{I}_3, \quad \text{and} \quad \frac{\langle a \rangle}{\langle a^{5} \rangle} \cong \frac{\mathbb{Z}}{\langle 5 \rangle} \cong \mathbb{I}_5. \quad ◂$$

Proposition 2.78. *If G is a finite abelian group and d is a divisor of $|G|$, then G contains a subgroup of order d.*

Proof. We prove the result by induction on $n = |G|$ for a prime divisor p of $|G|$. The base step $n = 1$ is true, for there are no prime divisors of 1. For the inductive step, choose $a \in G$ of order $k > 1$. If $p \mid k$, say $k = p\ell$, then Exercise 2.23 on page 62 says that a^ℓ has order p. If $p \nmid k$, consider the cyclic subgroup $H = \langle a \rangle$. Now $H \lhd G$, because G is abelian, and so the quotient group G/H exists. Note that $|G/H| = n/k$ is divisible by p, and so the inductive hypothesis gives an element $bH \in G/H$ of order p. If b has order m, then Exercise 2.47(i) on page 80 gives $p \mid m$. We have returned to the first case.

Let d be any divisor of $|G|$, and let p be a prime divisor of d. We have just seen that there is a subgroup $S \le G$ of order p. Now $S \lhd G$, because G is abelian, and G/S is a group of order n/p. By induction on $|G|$, G/S has a subgroup H^* of order d/p. The correspondence theorem gives $H^* = H/S$ for some subgroup H of G containing S, and $|H| = |H^*||S| = d$. •

Here is a construction of a new group from two given groups.

Definition. If H and K are groups, then their ***direct product***, denoted by $H \times K$, is the set of all ordered pairs (h, k) with $h \in H$ and $k \in K$ equipped with the operation

$$(h, k)(h', k') = (hh', kk').$$

It is easy to check that the direct product $H \times K$ is a group [the identity is $(1, 1)$ and $(h, k)^{-1} = (h^{-1}, k^{-1})$].

We now apply the first isomorphism theorem to direct products.

Proposition 2.79. *Let G and G' be groups, and let $K \lhd G$ and $K' \lhd G'$ be normal subgroups. Then $K \times K' \lhd G \times G'$, and there is an isomorphism*

$$(G \times G')/(K \times K') \cong (G/K) \times (G'/K').$$

Proof. Let $\pi : G \to G/K$ and $\pi' : G' \to G'/K'$ be the natural maps. It is routine to check that $f : G \times G' \to (G/K) \times (G'/K')$, given by

$$f : (g, g') \mapsto (\pi(g), \pi'(g')) = (gK, g'K')$$

is a surjective homomorphism with $\ker f = K \times K'$. The first isomorphism theorem now gives the desired isomorphism. •

Proposition 2.80. *If G is a group containing normal subgroups H and K with $H \cap K = \{1\}$ and $HK = G$, then $G \cong H \times K$.*

Proof. We show first that if $g \in G$, then the factorization $g = hk$, where $h \in H$ and $k \in K$, is unique. If $hk = h'k'$, then $h'^{-1}h = k'k^{-1} \in H \cap K = \{1\}$. Therefore, $h' = h$ and $k' = k$. We may now define a function $\varphi : G \to H \times K$ by $\varphi(g) = (h, k)$, where $g = hk$, $h \in H$, and $k \in K$. To see whether φ is a homomorphism, let $g' = h'k'$, so that $gg' = hkh'k'$. Hence, $\varphi(gg') = \varphi(hkh'k')$, which is not in the proper form for evaluation. If we knew that if $h \in H$ and $k \in K$, then $hk = kh$, then we could continue:

$$\begin{aligned} \varphi(hkh'k') &= \varphi(hh'kk') \\ &= (hh', kk') \\ &= (h, k)(h', k') \\ &= \varphi(g)\varphi(g'). \end{aligned}$$

Let $h \in H$ and $k \in K$. Since K is a normal subgroup, $(hkh^{-1})k^{-1} \in K$; since H is a normal subgroup, $h(kh^{-1}k^{-1}) \in H$. But $H \cap K = \{1\}$, so that $hkh^{-1}k^{-1} = 1$ and $hk = kh$. Finally, we show that the homomorphism φ is an isomorphism. If $(h, k) \in H \times K$, then the element $g \in G$ defined by $g = hk$ satisfies $\varphi(g) = (h, k)$; hence φ is surjective. If $\varphi(g) = (1, 1)$, then $g = 1$, so that $\ker \varphi = 1$ and φ is injective. Therefore, φ is an isomorphism. •

Remark. We must assume that both subgroups H and K are normal. For example, S_3 has subgroups $H = \langle(1\ 2\ 3)\rangle$ and $K = \langle(1\ 2)\rangle$. Now $H \lhd S_3$, $H \cap K = \{1\}$, and $HK = S_3$, but $S_3 \ncong H \times K$ (because the direct product is abelian). Of course, K is not a normal subgroup of S_3. ◂

Theorem 2.81. *If m and n are relatively prime, then*

$$\mathbb{I}_{mn} \cong \mathbb{I}_m \times \mathbb{I}_n.$$

Proof. If $a \in \mathbb{Z}$, denote its congruence class in $\mathbb{I}_m$ by $[a]_m$. The reader can show that the function $f : \mathbb{Z} \to \mathbb{I}_m \times \mathbb{I}_n$, given by $a \mapsto ([a]_m, [a]_n)$, is a homomorphism. We claim that $\ker f = \langle mn \rangle$. Clearly, $\langle mn \rangle \leq \ker f$. For the reverse inclusion, if $a \in \ker f$, then $[a]_m = [0]_m$ and $[a]_n = [0]_n$; that is, $a \equiv 0 \bmod m$ and $a \equiv 0 \bmod n$; that is, $m \mid a$ and $n \mid a$. Since m and n are relatively prime, $mn \mid a$, and so $a \in \langle mn \rangle$, that is, $\ker f \leq \langle mn \rangle$ and $\ker f = \langle mn \rangle$. The first isomorphism theorem now gives $\mathbb{Z}/\langle mn \rangle \cong \operatorname{im} f \leq \mathbb{I}_m \times \mathbb{I}_n$. But $\mathbb{Z}/\langle mn \rangle \cong \mathbb{I}_{mn}$ has mn elements, as does $\mathbb{I}_m \times \mathbb{I}_n$. We conclude that f is surjective. •

For example, it follows that $\mathbb{I}_6 \cong \mathbb{I}_2 \times \mathbb{I}_3$. Note that there is no isomorphism if m and n are not relatively prime. For example, $\mathbb{I}_4 \not\cong \mathbb{I}_2 \times \mathbb{I}_2$, for $\mathbb{I}_4$ has an element of order 4 and the direct product (which is isomorphic to the four-group $\mathbf{V}$) has no such element.

In light of Proposition 2.34, we may say that an element $a \in G$ has order n if $\langle a \rangle \cong \mathbb{I}_n$. Theorem 2.81 can now be interpreted as saying that if a and b are commuting elements having relatively prime orders m and n, then ab has order mn. Let us give a direct proof of this result.

Proposition 2.82. *Let G be a group, and let $a, b \in G$ be commuting elements of orders m and n, respectively. If $(m, n) = 1$, then ab has order mn.*

Proof. Since a and b commute, we have $(ab)^r = a^r b^r$ for all r, so that $(ab)^{mn} = a^{mn} b^{mn} = 1$. It suffices to prove that if $(ab)^k = 1$, then $mn \mid k$. If $1 = (ab)^k = a^k b^k$, then $a^k = b^{-k}$. Since a has order m, we have $1 = a^{mk} = b^{-mk}$. Since b has order n, Theorem 2.24 gives $n \mid mk$. As $(m, n) = 1$, however, Corollary 1.11 gives $n \mid k$; a similar argument gives $m \mid k$. Finally, Exercise 1.19 on page 13 shows that $mn \mid k$. Therefore, $mn \leq k$, and mn is the order of ab. •

Corollary 2.83. *If $(m, n) = 1$, then $\phi(mn) = \phi(m)\phi(n)$, where ϕ is the Euler ϕ-function.*

Proof. [15] Theorem 2.81 shows that the function $f : \mathbb{I}_{mn} \to \mathbb{I}_m \times \mathbb{I}_n$, given by $[a] \mapsto ([a]_m, [a]_n)$, is an isomorphism. The result will follow if we prove that $f(U(\mathbb{I}_{mn})) = U(\mathbb{I}_m) \times U(\mathbb{I}_n)$, for then

$$\begin{aligned} \phi(mn) &= |U(\mathbb{I}_{mn})| = |f(U(\mathbb{I}_{mn}))| \\ &= |U(\mathbb{I}_m) \times U(\mathbb{I}_n)| = |U(\mathbb{I}_m)| \cdot |U(\mathbb{I}_n)| = \phi(m)\phi(n). \end{aligned}$$

If $[a] \in U(\mathbb{I}_{mn})$, then $[a][b] = [1]$ for some $[b] \in \mathbb{I}_{mn}$, and

$$\begin{aligned} f([ab]) &= ([ab]_m, [ab]_n) = ([a]_m[b]_m, [a]_n[b]_n) \\ &= ([a]_m, [a]_n)([b]_m, [b]_n) = ([1]_m, [1]_n). \end{aligned}$$

Hence, $[1]_m = [a]_m[b]_m$ and $[1]_n = [a]_n[b]_n$, so that $f([a]) = ([a]_m, [a]_n) \in U(\mathbb{I}_m) \times U(\mathbb{I}_n)$, and $f(U(\mathbb{I}_{mn})) \subseteq U(\mathbb{I}_m) \times U(\mathbb{I}_n)$.

For the reverse inclusion, if $f([c]) = ([c]_m, [c]_n) \in U(\mathbb{I}_m) \times U(\mathbb{I}_n)$, then we must show that $[c] \in U(\mathbb{I}_{mn})$. There is $[d]_m \in \mathbb{I}_m$ with $[c]_m[d]_m = [1]_m$, and there is $[e]_n \in \mathbb{I}_n$ with

[15] See Exercise 3.50 on page 150 for a less cluttered proof.

$[c]_n[e]_n = [1]_n$. Since f is surjective, there is $b \in \mathbb{Z}$ with $([b]_m, [b]_n) = ([d]_m, [e]_n)$, so that

$$f([1]) = ([1]_m, [1]_n) = ([c]_m[b]_m, [c]_n[b]_n) = f([c][b]).$$

Since f is an injection, $[1] = [c][b]$ and $[c] \in U(\mathbb{I}_{mn})$. •

Corollary 2.84.

(i) *If p is a prime, then $\phi(p^e) = p^e - p^{e-1} = p^e\left(1 - \frac{1}{p}\right)$.*

(ii) *If $n = p_1^{e_1} \cdots p_t^{e_t}$ is the prime factorization of n, then*

$$\phi(n) = n\Big(1 - \frac{1}{p_1}\Big) \cdots \Big(1 - \frac{1}{p_t}\Big).$$

Sketch of Proof. Part (i) holds because $(k, p^e) = 1$ if and only if $p \nmid k$, while part (ii) follows from Corollary 2.83. •

Lemma 2.85. *A cyclic group of order n has a unique subgroup of order d, for each divisor d of n, and this subgroup is cyclic.*

Proof. Let $G = \langle a \rangle$. If $n = cd$, we show that a^c has order d (and so $\langle a^c \rangle$ is a subgroup of order d). Clearly $(a^c)^d = a^{cd} = a^n = 1$; we claim that d is the smallest such power. If $(a^c)^r = 1$, then $n \mid cr$ [Theorem 2.24]; hence $cr = ns = dcs$ for some integer s, and $r = ds \geq d$.

To prove uniqueness, assume that $\langle x \rangle$ is a subgroup of order d (recall that every subgroup of a cyclic group is cyclic, by Exercise 2.34 on page 72). Now $x = a^m$ and $1 = x^d = a^{md}$; hence $md = nk$ for some integer k. Therefore, $x = a^m = (a^{n/d})^k = (a^c)^k$, so that $\langle x \rangle \leq \langle a^c \rangle$. Since both subgroups have the same order d, it follows that $\langle x \rangle = \langle a^c \rangle$. •

Define an equivalence relation on a group G by $x \equiv y$ if $\langle x \rangle = \langle y \rangle$; that is, x and y are equivalent if they are generators of the same cyclic subgroup. Denote the equivalence class containing an element x by $\text{gen}(C)$, where $C = \langle x \rangle$; thus, $\text{gen}(C)$ consists of all the generators of C. As usual, equivalence classes form a partition, and so G is the disjoint union:

$$G = \bigcup_C \text{gen}(C),$$

where C ranges over all cyclic subgroups of G. In Theorem 2.33(ii), we proved that

$$|\text{gen}(C)| = \phi(n),$$

where ϕ is the Euler ϕ-function.

The next theorem will be used later to prove that the multiplicative group $\mathbb{I}_p^\times$ is cyclic.

Theorem 2.86. *A group G of order n is cyclic if and only if, for each divisor d of n, there is at most one cyclic subgroup of order d.*

Proof. If G is cyclic, then the result follows from Lemma 2.85. Conversely, write G as a disjoint union:

$$G = \bigcup_C \text{gen}(C).$$

Hence, $n = |G| = \sum |\text{gen}(C)|$, where the summation is over all cyclic subgroups C of G:

$$n = \sum_C |\text{gen}(C)| = \sum_C \phi(|C|).$$

By hypothesis, for any divisor d of n, the group G has at most one cyclic subgroup of order d. Therefore,

$$n = \sum_C |\text{gen}(C)| = \sum_C \phi(|C|) \le \sum_{d|n} \phi(d) = n,$$

the last equality being Corollary 1.39. Hence, for every divisor d of n, we must have $\phi(d)$ arising as $|\text{gen}(C)|$ for some cyclic subgroup C of G of order d. In particular, $\phi(n)$ arises; there is a cyclic subgroup of order n, and so G is cyclic. •

Here is a proof of the abelian case of the preceding theorem (shown to me by D. Leep).

Theorem. *If G is an abelian group of order n having at most one cyclic subgroup of order p for each prime divisor p of n, then G is cyclic.*

Proof. The proof is by induction on $n = |G|$, with the base step $n = 1$ obviously true. For the inductive step, note first that the hypothesis is inherited by subgroups of G. We claim that there is some element x in G whose order is a prime divisor p of $|G|$. Choose $y \in G$ with $y \neq 1$; its order k is a divisor of $|G|$, by Lagrange's theorem, and so $k = pm$ for some prime p. By Exercise 2.23 on page 62, the element $x = y^m$ has order p. Define $\theta \colon G \to G$ by $\theta : g \mapsto g^p$ (θ is a homomorphism because G is abelian). Now $x \in \ker\theta$, so that $|\ker\theta| \ge p$. If $|\ker\theta| > p$, then there would be more than p elements $g \in G$ satisfying $g^p = 1$, and this would force more than one subgroup of order p in G. Therefore, $|\ker\theta| = p$. By the first isomorphism theorem, $G/\ker\theta \cong \text{im}\,\theta \le G$. Thus, $\text{im}\,\theta$ is a subgroup of G of order n/p satisfying the inductive hypothesis, so there is an element $z \in \text{im}\,\theta$ with $\text{im}\,\theta = \langle z \rangle$. Moreover, since $z \in \text{im}\,\theta$, there is $b \in G$ with $z = b^p$. There are now two cases. If $p \nmid n/p$, then xz has order $p \cdot n/p = n$, by Proposition 2.82, and so $G = \langle xz \rangle$. If $p \mid n/p$, then Exercise 2.24 on page 62 shows that b has order n, and $G = \langle b \rangle$. •

EXERCISES

2.65 Prove that $U(\mathbb{I}_9) \cong \mathbb{I}_6$ and $U(\mathbb{I}_{15}) \cong \mathbb{I}_4 \times \mathbb{I}_2$.

2.66 (i) Let H and K be groups. Without using the first isomorphism theorem, prove that $H^* = \{(h, 1) : h \in H\}$ and $K^* = \{(1, k) : k \in K\}$ are normal subgroups of $H \times K$ with

$H \cong H^*$ and $K \cong K^*$, and $f\colon H \to (H \times K)/K^*$, defined by $f(h) = (h, 1)K^*$, is an isomorphism.

(ii) Use the first isomorphism theorem to prove that $K^* \lhd H \times K$ and that

$$(H \times K)/K^* \cong H.$$

Hint. Consider the function $f\colon H \times K \to H$ defined by $f\colon (h, k) \mapsto h$.

2.67 (i) Prove that $\mathrm{Aut}(\mathbf{V}) \cong S_3$ and that $\mathrm{Aut}(S_3) \cong S_3$. Conclude that nonisomorphic groups can have isomorphic automorphism groups.

(ii) Prove that $\mathrm{Aut}(\mathbb{Z}) \cong \mathbb{I}_2$. Conclude that an infinite group can have a finite automorphism group.

2.68 If G is a group for which $\mathrm{Aut}(G) = \{1\}$, prove that $|G| \le 2$.

2.69 Prove that if G is a group for which $G/Z(G)$ is cyclic, where $Z(G)$ denotes the center of G, then G is abelian.
Hint. If $G/Z(G)$ is cyclic, prove that a generator gives an element outside of $Z(G)$ which commutes with each element of G.

2.70 (i) Prove that $\mathbf{Q}/Z(\mathbf{Q}) \cong \mathbf{V}$, where $\mathbf{Q}$ is the group of quaternions and $\mathbf{V}$ is the four-group; conclude that the quotient of a group by its center can be abelian.

(ii) Prove that $\mathbf{Q}$ has no subgroup isomorphic to $\mathbf{V}$. Conclude that the quotient $\mathbf{Q}/Z(\mathbf{Q})$ is not isomorphic to a subgroup of $\mathbf{Q}$.

2.71 Let G be a finite group with $K \lhd G$. If $(|K|, [G : K]) = 1$, prove that K is the unique subgroup of G having order $|K|$.
Hint. If $H \le G$ and $|H| = |K|$, what happens to elements of H in G/K?

2.72 If H and K are subgroups of a group G, prove that HK is a subgroup of G if and only if $HK = KH$.
Hint. Use the fact that $H \subseteq HK$ and $K \subseteq HK$.

2.73 Let G be a group and regard $G \times G$ as the direct product of G with itself. If the multiplication $\mu\colon G \times G \to G$ is a group homomorphism, prove that G must be abelian.

2.74 Generalize Theorem 2.81 as follows. Let G be a finite (additive) abelian group of order mn, where $(m, n) = 1$. Define

$$G_m = \{g \in G : \text{order } (g) \mid m\} \text{ and } G_n = \{h \in G : \text{order } (h) \mid n\}.$$

(i) Prove that G_m and G_n are subgroups with $G_m \cap G_n = \{0\}$.

(ii) Prove that $G = G_m + G_n = \{g + h : g \in G_m \text{ and } h \in G_n\}$.

(iii) Prove that $G \cong G_m \times G_n$.

2.75 Let G be a finite group, let p be a prime, and let H be a normal subgroup of G. Prove that if both $|H|$ and $|G/H|$ are powers of p, then $|G|$ is a power of p.

2.76 If H and K are normal subgroups of a group G with $HK = G$, prove that

$$G/(H \cap K) \cong (G/H) \times (G/K).$$

Hint. If $\varphi\colon G \to (G/H) \times (G/K)$ is defined by $x \mapsto (xH, xK)$, then $\ker\varphi = H \cap K$; moreover, we have $G = HK$, so that

$$\bigcup_a aH = HK = \bigcup_b bK.$$

Definition. If $H_1, \ldots, H_n$ are groups, then their ***direct product***

$$H_1 \times \cdots \times H_n$$

is the set of all n-tuples $(h_1, \ldots, h_n)$, where $h_i \in H_i$ for all i, with coordinatewise multiplication:

$$(h_1, \ldots, h_n)(h'_1, \ldots, h'_n) = (h_1 h'_1, \ldots, h_n h'_n).$$

2.77 (i) Generalize Theorem 2.81 by proving that if the prime factorization of an integer m is $m = p_1^{e_1} \cdots p_n^{e_n}$, then

$$\mathbb{I}_m \cong \mathbb{I}_{p_1^{e_1}} \times \cdots \times \mathbb{I}_{p_n^{e_n}}.$$

(ii) Generalize Corollary 2.83 by proving that if the prime factorization of an integer m is $m = p_1^{e_1} \cdots p_n^{e_n}$, then

$$U(\mathbb{I}_m) \cong U(\mathbb{I}_{p_1^{e_1}}) \times \cdots \times U(\mathbb{I}_{p_n^{e_n}}).$$

2.7 Group Actions

Groups of permutations led us to abstract groups; the next result, due to A. Cayley, shows that abstract groups are not so far removed from permutations.

Theorem 2.87 (*Cayley*). *Every group G is isomorphic to a subgroup of the symmetric group S_G. In particular, if $|G| = n$, then G is isomorphic to a subgroup of S_n.*

Proof. For each $a \in G$, define "translation" $\tau_a \colon G \to G$ by $\tau_a(x) = ax$ for every $x \in G$ (if $a \neq 1$, then τ_a is not a homomorphism). For $a, b \in G$, $(\tau_a \circ \tau_b)(x) = \tau_a(\tau_b(x)) = \tau_a(bx) = a(bx) = (ab)x$, by associativity, so that

$$\tau_a \tau_b = \tau_{ab}.$$

It follows that each τ_a is a bijection, for its inverse is $\tau_{a^{-1}}$:

$$\tau_a \tau_{a^{-1}} = \tau_{aa^{-1}} = \tau_1 = 1_G = \tau_{a^{-1}a},$$

and so $\tau_a \in S_G$.

Define $\varphi : G \to S_G$ by $\varphi(a) = \tau_a$. Rewriting,

$$\varphi(a)\varphi(b) = \tau_a \tau_b = \tau_{ab} = \varphi(ab),$$

so that φ is a homomorphism. Finally, φ is an injection. If $\varphi(a) = \varphi(b)$, then $\tau_a = \tau_b$, and hence $\tau_a(x) = \tau_b(x)$ for all $x \in G$; in particular, when $x = 1$, this gives $a = b$, as desired.

The last statement follows from Exercise 2.39 on page 80, which says that if X is a set with $|X| = n$, then $S_X \cong S_n$. •

The reader may note, in the proof of Cayley's theorem, that the permutation τ_a is just the ath row of the multiplication table of G.

To tell the truth, Cayley's theorem itself is only mildly interesting. However, the identical proof works in a larger setting that is more interesting.

Theorem 2.88 (Representation on Cosets). *Let G be a group, and let H be a subgroup of G having finite index n. Then there exists a homomorphism $\varphi\colon G \to S_n$ with* $\ker\varphi \le H$.

Proof. Even though H may not be a normal subgroup, we still denote the family of all the cosets of H in G by G/H.

For each $a \in G$, define "translation" $\tau_a\colon G/H \to G/H$ by $\tau_a(xH) = axH$ for every $x \in G$. For $a, b \in G$,

$$(\tau_a \circ \tau_b)(xH) = \tau_a(\tau_b(xH)) = \tau_a(bxH) = a(bxH) = (ab)xH,$$

by associativity, so that

$$\tau_a\tau_b = \tau_{ab}.$$

It follows that each τ_a is a bijection, for its inverse is $\tau_{a^{-1}}$:

$$\tau_a\tau_{a^{-1}} = \tau_{aa^{-1}} = \tau_1 = 1_{G/H} = \tau_{a^{-1}}\tau_a,$$

and so $\tau_a \in S_{G/H}$. Define $\varphi : G \to S_{G/H}$ by $\varphi(a) = \tau_a$. Rewriting,

$$\varphi(a)\varphi(b) = \tau_a\tau_b = \tau_{ab} = \varphi(ab),$$

so that φ is a homomorphism. Finally, if $a \in \ker\varphi$, then $\varphi(a) = 1_{G/H}$, so that $\tau_a(xH) = xH$ for all $x \in G$; in particular, when $x = 1$, this gives $aH = H$, and $a \in H$, by Lemma 2.40(i). The result follows from Exercise 2.39 on page 80, for $|G/H| = n$, and so $S_{G/H} \cong S_n$. •

When $H = \{1\}$, this is the Cayley theorem.

We are now going to classify all groups of order up to 7. By Example 2.53, every group of prime order p is isomorphic to $\mathbb{I}_p$, and so, to isomorphism, there is just one group of order p. Of the possible orders through 7, four of them, 2, 3, 5, and 7, are primes, and so we need look only at orders 4 and 6.

Proposition 2.89. *Every group G of order 4 is isomorphic to either $\mathbb{I}_4$ or the four-group* $\mathbf{V}$. *Moreover, $\mathbb{I}_4$ and $\mathbf{V}$ are not isomorphic.*

Proof. By Lagrange's theorem, every element in G, other than 1, has order either 2 or 4. If there is an element of order 4, then G is cyclic. Otherwise, $x^2 = 1$ for all $x \in G$, so that Exercise 2.26 on page 62 shows that G is abelian.

If distinct elements x and y in G are chosen, neither being 1, then we quickly check that $xy \notin \{1, x, y\}$; hence,

$$G = \{1, x, y, xy\}.$$

It is easy to see that the bijection $f : G \to \mathbf{V}$, defined by $f(1) = 1$, $f(x) = (1\ 2)(3\ 4)$, $f(y) = (1\ 3)(2\ 4)$, and $f(xy) = (1\ 4)(2\ 3)$, is an isomorphism.

We have already seen, in Example 2.54, that $\mathbb{I}_4 \not\cong \mathbf{V}$. •

Proposition 2.90. *If G is a group of order 6, then G is isomorphic to either $\mathbb{I}_6$ or S_3. Moreover, $\mathbb{I}_6$ and S_3 are not isomorphic.*[16]

Proof. By Lagrange's theorem, the only possible orders of nonidentity elements are 2, 3, and 6. Of course, $G \cong \mathbb{I}_6$ if G has an element of order 6. Now Exercise 2.27 on page 62 shows that G must contain an element of order 2, say, t. We now consider the cases G abelian and G nonabelian separately.

Case 1. G is abelian.

If there is a second element of order 2, say, a, then it is easy to see, using $at = ta$, that $H = \{1, a, t, at\}$ is a subgroup of G. This contradicts Lagrange's theorem, because 4 is not a divisor of 6. It follows that G must contain an element b of order 3. But tb has order 6, by Proposition 2.82. Therefore, G is cyclic if it is abelian.

Case 2. G is not abelian.

If G has no elements of order 3, then $x^2 = 1$ for all $x \in G$, and G is abelian, by Exercise 2.26 on page 62. Therefore, G contains an element s of order 3 as well as the element t of order 2.

Now $|\langle s\rangle| = 3$, so that $[G : \langle s\rangle] = |G|/|\langle s\rangle| = 6/3 = 2$, and so $\langle s\rangle$ is a normal subgroup of G, by Proposition 2.62(ii). Since $t = t^{-1}$, we have $tst \in \langle s\rangle$; hence, $tst = s^i$ for $i = 0, 1$ or 2. Now $i \neq 0$, for $tst = s^0 = 1$ implies $s = 1$. If $i = 1$, then s and t commute, and this gives st of order 6, as in Case 1 (which forces G to be cyclic, hence abelian, contrary to our present hypothesis). Therefore, $tst = s^2 = s^{-1}$.

We now use Theorem 2.88 to construct an isomorphism $G \to S_3$. Let $H = \langle t\rangle$, and consider the homomorphism $\varphi : G \to S_{G/\langle t\rangle}$ given by

$$\varphi(g) : x\,\langle t\rangle \mapsto gx\,\langle t\rangle\,.$$

By the theorem, $\ker\varphi \leq \langle t\rangle$, so that either $\ker\varphi = \{1\}$ (and φ is injective), or $\ker\varphi = \langle t\rangle$. Now $G/\langle t\rangle = \{\langle t\rangle, s\,\langle t\rangle, s^2\,\langle t\rangle\}$, and, in two-rowed notation,

$$\varphi(t) = \begin{pmatrix} \langle t\rangle & s\,\langle t\rangle & s^2\,\langle t\rangle \\ t\,\langle t\rangle & ts\,\langle t\rangle & ts^2\,\langle t\rangle \end{pmatrix}.$$

If $\varphi(t)$ is the identity permutation, then $ts\,\langle t\rangle = s\,\langle t\rangle$, so that $s^{-1}ts \in \langle t\rangle = \{1, t\}$, by Lemma 2.40. But now $s^{-1}ts = t$ (it cannot be 1), hence $ts = st$, contradicting t and s not commuting. Therefore, $t \notin \ker\varphi$, and $\varphi\colon G \to S_{G/\langle t\rangle} \cong S_3$ is an injective homomorphism. Since both G and S_3 have order 6, φ must be a bijection, and so $G \cong S_3$.

It is clear that $\mathbb{I}_6$ and S_3 are not isomorphic, for one is abelian and the other is not. •

[16]Cayley states this proposition in an article he wrote in 1854. However, in 1878, in the *American Journal of Mathematics*, he wrote, "The general problem is to find all groups of a given order n; ... if $n = 6$, there are three groups; a group

$$1, \alpha, \alpha^2, \alpha^3, \alpha^4, \alpha^5 \quad (\alpha^6 = 1),$$

and two more groups

$$1, \beta, \beta^2, \alpha, \alpha\beta, \alpha\beta^2 \qquad (\alpha^2 = 1, \beta^3 = 1),$$

viz., in the first of these $\alpha\beta = \beta\alpha$ while in the other of them, we have $\alpha\beta = \beta^2\alpha$, $\alpha\beta^2 = \beta\alpha$." Cayley's list is $\mathbb{I}_6$, $\mathbb{I}_2 \times \mathbb{I}_3$, and S_3. Of course, $\mathbb{I}_2 \times \mathbb{I}_3 \cong \mathbb{I}_6$; even Homer nods.

One consequence of this result is another proof that $\mathbb{I}_6 \cong \mathbb{I}_2 \times \mathbb{I}_3$ (see Theorem 2.81).

Classifying groups of order 8 is more difficult, for we have not yet developed enough theory. It turns out that there are 5 nonisomorphic groups of order 8: Three are abelian: $\mathbb{I}_8$; $\mathbb{I}_4 \times \mathbb{I}_2$; $\mathbb{I}_2 \times \mathbb{I}_2 \times \mathbb{I}_2$; two are nonabelian: D_8; $\mathbf{Q}$.

We can continue this discussion for larger orders, but things soon get out of hand, as Table 2.4 shows. Making a telephone directory of groups is not the way to study them.

Order of Group	Number of Groups
2	1
4	2
8	5
16	14
32	51
64	267
128	2, 328
256	56, 092
512	10, 494, 213
1024	49, 487, 365, 422

Table 2.4.

Groups arose by abstracting the fundamental properties enjoyed by permutations. But there is an important feature of permutations that the axioms do not mention: Permutations are functions. We shall see that there are interesting consequences when this feature is restored.

Definition. If X is a set and G is a group, then G ***acts*** on X if there is a function $G \times X \to X$, denoted by $(g, x) \mapsto gx$, such that

(i) $(gh)x = g(hx)$ for all $g, h \in G$ and $x \in X$;

(ii) $1x = x$ for all $x \in X$, where 1 is the identity in G.

We also call X a ***G-set*** if G acts on X.

If a group G acts on a set X, then fixing the first variable, say g, gives a function $\alpha_g \colon X \to X$, namely, $\alpha_g \colon x \mapsto gx$. This function is a permutation of X, for its inverse is $\alpha_{g^{-1}}$:

$$\alpha_g \alpha_{g^{-1}} = \alpha_1 = 1_X = \alpha_{g^{-1}} \alpha_g.$$

It is easy to see that $\alpha \colon G \to S_X$, defined by $\alpha \colon g \mapsto \alpha_g$, is a homomorphism. Conversely, given any homomorphism $\varphi \colon G \to S_X$, define $gx = \varphi(g)(x)$. Thus, an action of a group G on a set X is another way of viewing a homomorphism $G \to S_X$.

Cayley's theorem says that a group G acts on itself by (left) translation, and its generalization, Theorem 2.88, shows that G also acts on the family of cosets of a subgroup H by (left) translation.

Example 2.91.
We show that G acts on itself by conjugation: that is, for each $g \in G$, define $\alpha_g \colon G \to G$ to be conjugation

$$\alpha_g(x) = gxg^{-1}.$$

To verify axiom (i), note that for each $x \in G$,

$$\begin{aligned}(\alpha_g \circ \alpha_h)(x) &= \alpha_g(\alpha_h(x)) \\ &= \alpha_g(hxh^{-1}) \\ &= g(hxh^{-1})g^{-1} \\ &= (gh)x(gh)^{-1} \\ &= \alpha_{gh}(x).\end{aligned}$$

Therefore, $\alpha_g \circ \alpha_h = \alpha_{gh}$.

To prove axiom (ii), note that for each $x \in G$,

$$\alpha_1(x) = 1x1^{-1} = x,$$

and so $\alpha_1 = 1_G$. ◄

The following two definitions are fundamental.

Definition. If G acts on X and $x \in X$, then the ***orbit*** of x, denoted by $\mathcal{O}(x)$, is the subset of X

$$\mathcal{O}(x) = \{gx : g \in G\} \subseteq X;$$

the ***stabilizer*** of x, denoted by G_x, is the subgroup

$$G_x = \{g \in G : gx = x\} \leq G.$$

If G acts on a set X, define a relation on X by $x \equiv y$ in case there exists $g \in G$ with $y = gx$. It is easy to see that this is an equivalence relation whose equivalence classes are the orbits.

Let us find some orbits and stabilizers.

Example 2.92.
(i) Cayley's theorem says that G acts on itself by translations: $\tau_g \colon a \mapsto ga$. If $a \in G$, then the orbit $\mathcal{O}(a) = G$, for if $b \in G$, then $b = (ba^{-1})a = \tau_{ba^{-1}}(a)$. The stabilizer G_a of $a \in G$ is $\{1\}$, for if $a = \tau_g(a) = ga$, then $g = 1$. We say that G acts ***transitively*** on X if there is only one orbit.

(ii) When G acts on G/H (the family of cosets of a subgroup H) by translations τ_g : $aH \mapsto gaH$, then the orbit $\mathcal{O}(aH) = G/H$, for if $bH \in G/H$, then $\tau_{ba^{-1}} \colon aH \mapsto bH$. Thus, G acts transitively on G/H. The stabilizer G_{aH} of aH is aHa^{-1}, for $gaH = aH$ if and only if $a^{-1}ga \in H$ if and only if $g \in aHa^{-1}$. ◄

Example 2.93.
When a group G acts on itself by conjugation, then the orbit $\mathcal{O}(x)$ is

$$\{y \in G : y = axa^{-1} \text{ for some } a \in G\};$$

in this case, $\mathcal{O}(x)$ is called the ***conjugacy class*** of x, and it is commonly denoted by x^G. For example, Theorem 2.9 shows that if $\alpha \in S_n$, then the conjugacy class of α consists of all the permutations in S_n having the same cycle structure as α. As a second example, an element z lies in the center $Z(G)$ if and only if $z^G = \{z\}$; that is, no other elements in G are conjugate to z.

If $x \in G$, then the stabilizer G_x of x is

$$C_G(x) = \{g \in G : gxg^{-1} = x\}.$$

This subgroup of G, consisting of all $g \in G$ that commute with x, is called the ***centralizer*** of x in G. ◀

Example 2.94.
Every group G acts on the set X of all its subgroups, by conjugation: If $a \in G$, then a acts by $H \mapsto aHa^{-1}$, where $H \le G$.

If H is a subgroup of a group G, then a ***conjugate*** of H is a subgroup of G of the form

$$aHa^{-1} = \{aha^{-1} : h \in H\},$$

where $a \in G$.

Since conjugation $h \mapsto aha^{-1}$ is an injection $H \to G$ with image aHa^{-1}, it follows that conjugate subgroups of G are isomorphic. For example, in S_3, all cyclic subgroups of order 2 are conjugate (for their generators are conjugate).

The orbit of a subgroup H consists of all its conjugates; notice that H is the only element in its orbit if and only if $H \lhd G$; that is, $aHa^{-1} = H$ for all $a \in G$. The stabilizer of H is

$$N_G(H) = \{g \in G : gHg^{-1} = H\}.$$

This subgroup of G is called the ***normalizer*** of H in G. ◀

Example 2.95.
Let $X =$ the vertices $\{v_1, v_2, v_3, v_4\}$ of a square, and let G be the dihedral group D_8 acting on X, as in Figure 2.8 on page 102 (for clarity, the vertices in the figure are labeled 1, 2, 3, 4 instead of v_1, v_2, v_3, v_4).

$$\begin{aligned} G = \{&\text{rotations} : (1), (1\ 2\ 3\ 4), (1\ 3)(2\ 4), (1\ 4\ 3\ 2); \\ &\text{reflections} : (2\ 4), (1\ 3), (1\ 2)(3\ 4), (1\ 4)(2\ 3)\}. \end{aligned}$$

For each vertex $v_i \in X$, there is some $g \in G$ with $gv_1 = v_i$; therefore, $\mathcal{O}(v_1) = X$ and D_8 acts transitively.

What is the stabilizer G_{v_1} of v_1? Aside from the identity, there is only one $g \in D_8$ fixing v_1, namely, $g = (2\ 4)$; therefore G_{v_1} is a subgroup of order 2. (This example can be generalized to the dihedral group D_{2n} acting on a regular n-gon.) ◀

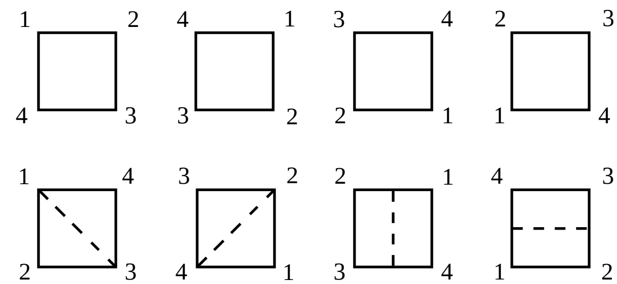


Figure 2.8

Example 2.96.
Let $X = \{1, 2, \ldots, n\}$, let $\alpha \in S_n$, and regard the cyclic group $G = \langle\alpha\rangle$ as acting on X. If $i \in X$, then

$$\mathcal{O}(i) = \{\alpha^k(i) : k \in \mathbb{Z}\}.$$

Let the complete factorization of α be $\alpha = \beta_1 \cdots \beta_{t(\alpha)}$, and let $i = i_1$ be moved by α. If the cycle involving i_1 is $\beta_j = (i_1\ i_2\ \ldots\ i_r)$, then the proof of Theorem 2.3 shows that $i_{k+1} = \alpha^k(i_1)$ for all $k < r$. Therefore,

$$\mathcal{O}(i) = \{i_1, i_2, \ldots, i_r\},$$

where $i = i_1$. It follows that $|\mathcal{O}(i)| = r$. The stabilizer G_ℓ of a number ℓ is G if α fixes ℓ; however, if α moves ℓ, then G_ℓ depends on the size of the orbit $\mathcal{O}(\ell)$. For example, if $\alpha = (1\ 2\ 3)(4\ 5)(6)$, then $G_6 = G$, $G_1 = \langle\alpha^3\rangle$, and $G_4 = \langle\alpha^2\rangle$. ◄

Proposition 2.97. *If G acts on a set X, then X is the disjoint union of the orbits. If X is finite, then*

$$|X| = \sum_i |\mathcal{O}(x_i)|,$$

where one x_i is chosen from each orbit.

Proof. As we have mentioned earlier, the relation on X, given by $x \equiv y$ if there exists $g \in G$ with $y = gx$, is an equivalence relation whose equivalence classes are the orbits. Therefore, the orbits partition X.

The count given in the second statement is correct: Since the orbits are disjoint, no element in X is counted twice. •

Here is the connection between orbits and stabilizers.

Theorem 2.98. *If G acts on a set X and $x \in X$, then*

$$|\mathcal{O}(x)| = [G : G_x]$$

the index of the stabilizer G_x in G.

Proof. Let G/G_x denote the family of all the left cosets of G_x in G. We will exhibit a bijection $\varphi\colon G/G_x \to \mathcal{O}(x)$, and this will give the result, since $|G/G_x| = [G : G_x]$. Define $\varphi\colon gG_x \mapsto gx$. Now φ is well-defined: If $gG_x = hG_x$, then $h = gf$ for some $f \in G_x$; that is, $fx = x$; hence, $hx = gfx = gx$. Now φ is an injection: if $gx = \varphi(gG_x) = \varphi(hG_x) = hx$, then $h^{-1}gx = x$; hence, $h^{-1}g \in G_x$, and $gG_x = hG_x$. Lastly, φ is a surjection: if $y \in \mathcal{O}(x)$, then $y = gx$ for some $g \in G$, and so $y = \varphi(gG_x)$. •

In Example 2.95, D_8 acting on the four corners of a square, we saw that $|\mathcal{O}(v_1)| = 4$, $|G_{v_1}| = 2$, and $[G : G_{v_1}] = 8/2 = 4$. In Example 2.96, $G = \langle\alpha\rangle \leq S_n$ acting on $X = \{1, 2, \ldots, n\}$, we saw that if, in the complete factorization of α into disjoint cycles $\alpha = \beta_1 \cdots \beta_{t(\alpha)}$, the r-cycle β_j moves ℓ, then $r = |\mathcal{O}(\ell)|$ for any ℓ occurring in β_j. Theorem 2.98 says that r is a divisor of the order k of α. (But Theorem 2.25 tells us more: k is the lcm of the lengths of the cycles occurring in the factorization.)

Corollary 2.99. *If a finite group G acts on a set X, then the number of elements in any orbit is a divisor of $|G|$.*

Proof. This follows at once from Lagrange's theorem. •

In Example 2.5(i), there is a table displaying the number of permutations in S_4 of each cycle structure; these numbers are 1, 6, 8, 6, 3. Note that each of these numbers is a divisor of $|S_4| = 24$, In Example 2.5(ii), we saw that the corresponding numbers are 1, 10, 20, 30, 24, 20, and 15, and these are all divisors of $|S_5| = 120$. We now recognize these subsets as being conjugacy classes, and the next corollary explains why these numbers divide the group order.

Corollary 2.100. *If x lies in a finite group G, then the number of conjugates of x is the index of its centralizer:*

$$|x^G| = [G : C_G(x)],$$

and hence it is a divisor of $|G|$.

Proof. As in Example 2.93, the orbit of x is its conjugacy class x^G, and the stabilizer G_x is the centralizer $C_G(x)$. •

Proposition 2.101. *If H is a subgroup of a finite group G, then the number of conjugates of H in G is $[G : N_G(H)]$.*

Proof. As in Example 2.94, the orbit of H is the family of all its conjugates, and the stabilizer is its normalizer $N_G(H)$. •

There are some interesting applications of group actions to counting problems, which we will give at the end of this section. Let us first apply group actions to group theory.

When we began classifying groups of order 6, it would have been helpful to be able to assert that any such group has an element of order 3 (we were able to use an earlier exercise to assert the existence of an element of order 2). We now prove that if p is a prime divisor of $|G|$, where G is a finite group G, then G contains an element of order p.

Theorem 2.102 (***Cauchy***). *If G is a finite group whose order is divisible by a prime p, then G contains an element of order p.*

Proof. We prove the theorem by induction on $m \geq 1$, where $|G| = pm$. The base step $m = 1$ is true, for Lagrange's theorem shows that every nonidentity element in a group of order p has order p.

Let us now prove the inductive step. If $x \in G$, then the number of conjugates of x is $|x^G| = [G : C_G(x)]$, where $C_G(x)$ is the centralizer of x in G. As noted earlier, if $x \notin Z(G)$, then x^G has more than one element, and so $|C_G(x)| < |G|$. If $p \mid |C_G(x)|$ for some noncentral x, then the inductive hypothesis says there is an element of order p in $C_G(x) \leq G$, and we are done. Therefore, we may assume that $p \nmid |C_G(x)|$ for all noncentral $x \in G$. Better, since p is a prime and $|G| = [G : C_G(x)]|C_G(x)|$, Euclid's lemma gives

$$p \mid [G : C_G(x)].$$

After recalling that $Z(G)$ consists of all those elements $x \in G$ with $|x^G| = 1$, we may use Proposition 2.97 to see

$$|G| = |Z(G)| + \sum_i [G : C_G(x_i)],$$

where one x_i is selected from each conjugacy class having more than one element. Since $|G|$ and all $[G : C_G(x_i)]$ are divisible by p, it follows that $|Z(G)|$ is divisible by p. But $Z(G)$ is abelian, and so Proposition 2.78 says that $Z(G)$, and hence G, contains an element of order p. •

Definition. The ***class equation*** of a finite group G is

$$|G| = |Z(G)| + \sum_i [G : C_G(x_i)],$$

where one x_i is selected from each conjugacy class having more than one element.

Definition. If p is a prime, then a finite group G is called a ***p-group*** if $|G| = p^n$ for some $n \geq 0$. (See Exercise 2.81 on page 112 for the definition of an infinite p-group.)

We have seen examples of groups whose center is trivial; for example, $Z(S_3) = \{1\}$. For p-groups, however, this is never true.

Theorem 2.103. *If p is a prime and G is a p-group, then $Z(G) \neq \{1\}$.*

Proof. Consider the class equation

$$|G| = |Z(G)| + \sum_i [G : C_G(x_i)].$$

Each $C_G(x_i)$ is a proper subgroup of G, for $x_i \notin Z(G)$. Since G is a p-group, $[G : C_G(x_i)]$ is a divisor of $|G|$, hence is itself a power of p. Thus, p divides each of the terms in the class equation other than $|Z(G)|$, and so $p \mid |Z(G)|$ as well. Therefore, $Z(G) \neq \{1\}$. •

Corollary 2.104. *If p is a prime, then every group G of order p^2 is abelian.*

Proof. If G is not abelian, then its center $Z(G)$ is a proper subgroup, so that $|Z(G)| = 1$ or p, by Lagrange's theorem. But Theorem 2.103 says that $Z(G) \neq \{1\}$, and so $|Z(G)| = p$. The center is always a normal subgroup, so that the quotient $G/Z(G)$ is defined; it has order p, and hence $G/Z(G)$ is cyclic. This contradicts Exercise 2.69 on page 95. •

Example 2.105.
Who would have guessed that Cauchy's theorem (if G is a group whose order is a multiple of a prime p, then G has an element of order p) and Fermat's theorem (if p is prime, then $a^p \equiv a \bmod p$) are special cases of some common theorem? The elementary yet ingenious proof of Cauchy's theorem is due to J. H. McKay in 1959 (see Montgomery and Ralston, *Selected Papers in Algebra*); A. Mann showed me that McKay's argument also proves Fermat's theorem. If G is a finite group and p is a prime, denote the cartesian product of p copies of G by G^p, and define

$$X = \{(a_0, a_1, \ldots, a_{p-1}) \in G^p : a_0 a_1 \ldots a_{p-1} = 1\}.$$

Note that $|X| = |G|^{p-1}$, for having chosen the last $p - 1$ entries arbitrarily, the 0th entry must equal $(a_1 a_2 \cdots a_{p-1})^{-1}$. Introduce an action of $\mathbb{I}_p$ on X by defining, for $0 \leq i \leq p - 1$,

$$[i](a_0, a_1, \ldots, a_{p-1}) = (a_i, a_{i+1}, \ldots, a_{p-1}, a_0, a_1, \ldots, a_i).$$

The product of the entries in the new p-tuple is a conjugate of $a_0 a_1 \cdots a_{p-1}$:

$$a_i a_{i+1} \cdots a_{p-1} a_0 a_1 \cdots a_i = (a_0 a_1 \cdots a_i)^{-1}(a_0 a_1 \cdots a_{p-1})(a_0 a_1 \cdots a_i).$$

This conjugate is 1 (for $g^{-1}1g = 1$), and so $[i](a_0, a_1, \ldots, a_{p-1}) \in X$. By Corollary 2.99, the size of every orbit of X is a divisor of $|\mathbb{I}_p| = p$; since p is prime, these sizes are either 1 or p. Now orbits with just one element consist of a p-tuple all of whose entries a_i are equal, for all cyclic permutations of the p-tuple are the same. In other words, such an orbit corresponds to an element $a \in G$ with $a^p = 1$. Clearly, $(1, 1, \ldots, 1)$ is such an orbit; if it were the only such, then we would have

$$|G|^{p-1} = |X| = 1 + kp$$

for some $k \geq 0$; that is, $|G|^{p-1} \equiv 1 \bmod p$. If p is a divisor of $|G|$, then we have a contradiction, for $|G|^{p-1} \equiv 0 \bmod p$. We have thus proved Cauchy's theorem: If a prime p is a divisor of $|G|$, then G has an element of order p.

Suppose now that G is a group of order n, say, $G = \mathbb{I}_n$, and that p is not a divisor of n. By Lagrange's theorem, G has no elements of order p, so that if $a^p = 1$, then $a = 1$. Therefore, the only orbit in G^p of size 1 is $(1, 1, \ldots, 1)$, and so

$$n^{p-1} = |G|^{p-1} = |X| = 1 + kp;$$

that is, if p is not a divisor of n, then $n^{p-1} \equiv 1 \bmod p$. Multiplying both sides by n, we have $n^p \equiv n \bmod p$, a congruence also holding when p is a divisor of n; this is Fermat's theorem. ◂

We have seen, in Proposition 2.64, that A_4 is a group of order 12 having no subgroup of order 6. Thus, the assertion that if d is a divisor of $|G|$, then G must have a subgroup of order d, is false. However, this assertion is true when G is a p-group.

Proposition 2.106. *If G is a group of order $|G| = p^e$, then G has a normal subgroup of order p^k for every $k \le e$.*

Proof. We prove the result by induction on $e \ge 0$. The base step is obviously true, and so we proceed to the inductive step. By Theorem 2.103, the center of G is a nontrivial normal subgroup: $Z(G) \ne \{1\}$. Let $Z \le Z(G)$ be a subgroup of order p; as any subgroup of $Z(G)$, the subgroup Z is a normal subgroup of G. If $k \le e$, then $p^{k-1} \le p^{e-1} = |G/Z|$. By induction, G/Z has a normal subgroup H^* of order p^{k-1}. The correspondence theorem says there is a subgroup H of G containing Z with $H^* = H/Z$; moreover, $H^* \lhd G/Z$ implies $H \lhd G$. But $|H/Z| = p^{k-1}$ implies $|H| = p^k$, as desired. •

Abelian groups (and the quaternions) have the property that every subgroup is normal. At the opposite pole are groups having no normal subgroups other than the two obvious ones: $\{1\}$ and G.

Definition. A group $G \ne \{1\}$ is called ***simple*** if G has no normal subgroups other than $\{1\}$ and G itself.

Proposition 2.107. *An abelian group G is simple if and only if it is finite and of prime order.*

Proof. If G is finite of prime order p, then G has no subgroups H other than $\{1\}$ and G, otherwise Lagrange's theorem would show that $|H|$ is a divisor of p. Therefore, G is simple.

Conversely, assume that G is simple. Since G is abelian, every subgroup is normal, and so G has no subgroups other than $\{1\}$ and G. Choose $x \in G$ with $x \ne 1$. Since $\langle x \rangle$ is a subgroup, we have $\langle x \rangle = G$. If x has infinite order, then all the powers of x are distinct, and so $\langle x^2 \rangle < \langle x \rangle$ is a forbidden subgroup of $\langle x \rangle$, a contradiction. Therefore, every $x \in G$ has finite order. If x has (finite) order m and if m is composite, say $m = k\ell$, then $\langle x^k \rangle$ is a proper nontrivial subgroup of $\langle x \rangle$, a contradiction. Therefore, $G = \langle x \rangle$ has prime order. •

We are now going to show that A_5 is a nonabelian simple group (indeed, it is the smallest such; there is no nonabelian simple group of order less than 60).

Suppose that an element $x \in G$ has k conjugates; that is

$$|x^G| = |\{gxg^{-1} : g \in G\}| = k.$$

If there is a subgroup $H \le G$ with $x \in H \le G$, how many conjugates does x have in H? Since

$$x^H = \{hxh^{-1} : h \in H\} \subseteq \{gxg^{-1} : g \in G\} = x^G,$$

we have $|x^H| \leq |x^G|$. It is possible that there is strict inequality $|x^H| < |x^G|$. For example, take $G = S_3$, $x = (1\ 2)$, and $H = \langle x \rangle$. We know that $|x^G| = 3$ (because all transpositions are conjugate), whereas $|x^H| = 1$ (because H is abelian).

Now let us consider this question, in particular, for $G = S_5$, $x = (1\ 2\ 3)$, and $H = A_5$.

Lemma 2.108. *All* 3*-cycles are conjugate in* A_5.

Proof. Let $G = S_5$, $\alpha = (1\ 2\ 3)$, and $H = A_5$. We know that $|\alpha^{S_5}| = 20$, for there are twenty 3-cycles in S_5, as we saw in Example 2.5(ii). Therefore, $20 = |S_5|/|C_{S_5}(\alpha)| = 120/|C_{S_5}(\alpha)|$, by Corollary 2.100, so that $|C_{S_5}(\alpha)| = 6$; that is, there are exactly six permutations in S_5 that commute with α. Here they are:

$$(1),\ (1\ 2\ 3),\ (1\ 3\ 2),\ (4\ 5),\ (4\ 5)(1\ 2\ 3),\ (4\ 5)(1\ 3\ 2).$$

The last three of these are odd permutations, so that $|C_{A_5}(\alpha)| = 3$. We conclude that

$$|\alpha^{A_5}| = |A_5|/|C_{A_5}(\alpha)| = 60/3 = 20;$$

that is, all 3-cycles are conjugate to $\alpha = (1\ 2\ 3)$ in A_5. •

This lemma can be generalized from A_5 to all A_n for $n \geq 5$; see Exercise 2.91 on page 113.

Lemma 2.109. *If* $n \geq 3$, *every element in* A_n *is a* 3*-cycle or a product of* 3*-cycles.*

Proof. If $\alpha \in A_n$, then α is a product of an even number of transpositions:

$$\alpha = \tau_1 \tau_2 \cdots \tau_{2q-1} \tau_{2q}.$$

Of course, we may assume that adjacent τ's are distinct. As the transpositions may be grouped in pairs $\tau_{2i-1}\tau_{2i}$, it suffices to consider products $\tau\tau'$, where τ and τ' are transpositions. If τ and τ' are not disjoint, then $\tau = (i\ j)$, $\tau' = (i\ k)$, and $\tau\tau' = (i\ k\ j)$; if τ and τ' are disjoint, then $\tau\tau' = (i\ j)(k\ \ell) = (i\ j)(j\ k)(j\ k)(k\ \ell) = (i\ j\ k)(j\ k\ \ell)$. •

Theorem 2.110. A_5 *is a simple group.*

Proof. We shall show that if H is a normal subgroup of A_5 and $H \neq \{(1)\}$, then $H = A_5$. Now if H contains a 3-cycle, then normality forces H to contain all its conjugates. By Lemma 2.108, H contains every 3-cycle, and by Lemma 2.109, $H = A_5$. Therefore, it suffices to prove that H contains a 3-cycle.

As $H \neq \{(1)\}$, it contains some $\sigma \neq (1)$. We may assume, after a harmless relabeling, that either $\sigma = (1\ 2\ 3)$, $\sigma = (1\ 2)(3\ 4)$, or $\sigma = (1\ 2\ 3\ 4\ 5)$. As we have just remarked, we are done if σ is a 3-cycle.

If $\sigma = (1\ 2)(3\ 4)$, define $\tau = (1\ 2)(3\ 5)$. Now H contains $(\tau\sigma\tau^{-1})\sigma^{-1}$, because it is a normal subgroup, and $\tau\sigma\tau^{-1}\sigma^{-1} = (3\ 5\ 4)$, as the reader should check. If $\sigma = (1\ 2\ 3\ 4\ 5)$, define $\rho = (1\ 3\ 2)$; now H contains $\rho\sigma\rho^{-1}\sigma^{-1} = (1\ 3\ 4)$, as the reader should also check.

We have shown, in all cases, that H contains a 3-cycle. Therefore, the only normal subgroups in A_5 are $\{(1)\}$ and A_5 itself, and so A_5 is simple. •

Theorem 2.110 turns out to be the basic reason why the quadratic formula has no generalization giving the roots of polynomials of degree 5 or higher (see Theorem 4.27).

Without much more effort, we can prove that the alternating groups A_n are simple for all $n \geq 5$. Observe that A_4 is not simple, for the four-group $\mathbf{V}$ is a normal subgroup of A_4.

Lemma 2.111. *A_6 is a simple group.*

Proof. Let $H \neq \{(1)\}$ be a normal subgroup of A_6; we must show that $H = A_6$. Assume that there is some $\alpha \in H$ with $\alpha \neq (1)$ that fixes some i, where $1 \leq i \leq 6$. Define

$$F = \{\sigma \in A_6 : \sigma(i) = i\}.$$

Note that $\alpha \in H \cap F$, so that $H \cap F \neq \{(1)\}$. The second isomorphism theorem gives $H \cap F \lhd F$. But F is simple, for $F \cong A_5$, and so the only normal subgroups in F are $\{(1)\}$ and F. Since $H \cap F \neq \{(1)\}$, we have $H \cap F = F$; that is, $F \leq H$. It follows that H contains a 3-cycle, and so $H = A_6$, by Exercise 2.91 on page 113.

We may now assume that there is no $\alpha \in H$ with $\alpha \neq (1)$ that fixes some i with $1 \leq i \leq 6$. If we consider the cycle structures of permutations in A_6, however, any such α must have cycle structure $(1\ 2)(3\ 4\ 5\ 6)$ or $(1\ 2\ 3)(4\ 5\ 6)$. In the first case, $\alpha^2 \in H$ is a nontrivial permutation that fixes 1 (and also 2), a contradiction. In the second case, H contains $\alpha(\beta\alpha^{-1}\beta^{-1})$, where $\beta = (2\ 3\ 4)$, and it is easily checked that this is a nontrivial element in H which fixes 1, another contradiction. Therefore, no such normal subgroup H can exist, and so A_6 is a simple group. •

Theorem 2.112. *A_n is a simple group for all $n \geq 5$.*

Proof. If H is a nontrivial normal subgroup of A_n, that is, $H \neq \{(1)\}$, then we must show that $H = A_n$; by Exercise 2.91 on page 113, it suffices to prove that H contains a 3-cycle. If $\beta \in H$ is nontrivial, then there exists some i that β moves; say, $\beta(i) = j \neq i$. Choose a 3-cycle α that fixes i and moves j. The permutations α and β do not commute: $\beta\alpha(i) = \beta(i) = j$, while $\alpha\beta(i) = \alpha(j) \neq j$. It follows that $\gamma = (\alpha\beta\alpha^{-1})\beta^{-1}$ is a nontrivial element of H. But $\beta\alpha^{-1}\beta^{-1}$ is a 3-cycle, by Theorem 2.9, and so $\gamma = \alpha(\beta\alpha^{-1}\beta^{-1})$ is a product of two 3-cycles. Hence, γ moves at most 6 symbols, say, $i_1, \ldots, i_6$ (if γ moves fewer than 6 symbols, just adjoin others so we have a list of 6). Define

$$F = \{\sigma \in A_n : \sigma \text{ fixes all } i \neq i_1, \ldots, i_6\}.$$

Now $F \cong A_6$ and $\gamma \in H \cap F$. Hence, $H \cap F$ is a nontrivial normal subgroup of F. But F is simple, being isomorphic to A_6, and so $H \cap F = F$; that is, $F \leq H$. Therefore, H contains a 3-cycle, and so $H = A_n$; the proof is complete. •

We now use groups to solve some difficult counting problems.

Theorem 2.113 (Burnside's Lemma[17]). *Let G act on a finite set X. If N is the number of orbits, then*

$$N = \frac{1}{|G|} \sum_{\tau \in G} \mathrm{Fix}(\tau),$$

where $\mathrm{Fix}(\tau)$ *is the number of* $x \in X$ *fixed by* τ.

Proof. List the elements of X as follows: Choose $x_1 \in X$, and then list all the elements $x_1, x_2, \ldots, x_r$ in the orbit $\mathcal{O}(x_1)$; then choose $x_{r+1} \notin \mathcal{O}(x_1)$, and list the elements $x_{r+1}, x_{r+2}, \ldots$ in $\mathcal{O}(x_{r+1})$; continue this procedure until all the elements of X are listed. Now list the elements $\tau_1, \tau_2, \ldots, \tau_n$ of G, and form the following array, where

$$f_{i,j} = \begin{cases} 1 & \text{if } \tau_i \text{ fixes } x_j \\ 0 & \text{if } \tau_i \text{ moves } x_j. \end{cases}$$

	x_1	x_2	$\cdots$	x_{r+1}	x_{r+2}	$\cdots$
τ_1	$f_{1,1}$	$f_{1,2}$	$\cdots$	$f_{1,r+1}$	$f_{1,r+2}$	$\cdots$
τ_2	$f_{2,1}$	$f_{2,2}$	$\cdots$	$f_{2,r+1}$	$f_{2,r+2}$	$\cdots$
τ_i	$f_{i,1}$	$f_{i,2}$	$\cdots$	$f_{i,r+1}$	$f_{i,r+2}$	$\cdots$
τ_n	$f_{n,1}$	$f_{n,2}$	$\cdots$	$f_{n,r+1}$	$f_{n,r+2}$	$\cdots$

Now $\mathrm{Fix}(\tau_i)$, the number of x fixed by τ_i, is the number of 1's in the ith row of the array; therefore, $\sum_{\tau \in G} \mathrm{Fix}(\tau)$ is the total number of 1's in the array. Let us now look at the columns. The number of 1's in the first column is the number of τ_i that fix x_1; by definition, these τ_i comprise G_{x_1}. Thus, the number of 1's in column 1 is $|G_{x_1}|$. Similarly, the number of 1's in column 2 is $|G_{x_2}|$. By Exercise 2.99 on page 114, $|G_{x_1}| = |G_{x_2}|$. By Theorem 2.98, the number of 1's in the r columns labeled by the $x_i \in \mathcal{O}(x_1)$ is thus

$$r|G_{x_1}| = |\mathcal{O}(x_1)| \cdot |G_{x_1}| = \big(|G|/|G_{x_1}|\big)\, |G_{x_1}| = |G|.$$

The same is true for any other orbit: Its columns contain exactly $|G|$ 1's. Therefore, if there are N orbits, there are $N|G|$ 1's in the array. We conclude that

$$\sum_{\tau \in G} \mathrm{Fix}(\tau) = N|G|. \quad \bullet$$

We are going to use Burnside's lemma to solve problems of the following sort. How many striped flags are there having six stripes (of equal width) each of which can be colored red, white, or blue? Clearly, the two flags in Figure 2.9 are the same: The bottom flag is just the top one turned over.

[17]Burnside himself attributed this lemma to F. G. Frobenius. To avoid the confusion that would be caused by changing a popular name, P. M. Neumann has suggested that it be called "not-Burnside's lemma." W. Burnside was a fine mathematician, and there do exist theorems properly attributed to him. For example, Burnside proved that if p and q are primes, then there are no simple groups of order $p^m q^n$.

r	w	b	r	w	b

b	w	r	b	w	r

Figure 2.9

Let X be the set of all 6-tuples of colors; if $x \in X$, then

$$x = (c_1, c_2, c_3, c_4, c_5, c_6),$$

where each c_i denotes either red, white, or blue. Let τ be the permutation that reverses all the indices:

$$\tau = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 6 & 5 & 4 & 3 & 2 & 1 \end{pmatrix} = (1\ 6)(2\ 5)(3\ 4)$$

(thus, τ "turns over" each 6-tuple x of colored stripes). The cyclic group $G = \langle \tau \rangle$ acts on X; since $|G| = 2$, the orbit of any 6-tuple x consists of either 1 or 2 elements: Either τ fixes x or it does not. Since a flag is unchanged by turning it over, it is reasonable to identify a flag with an orbit of a 6-tuple. For example, the orbit consisting of the 6-tuples

$$(r, w, b, r, w, b) \quad \text{and} \quad (b, w, r, b, w, r)$$

describes the flag in Figure 2.9. The number of flags is thus the number N of orbits; by Burnside's lemma, $N = \frac{1}{2}[\text{Fix}((1)) + \text{Fix}(\tau)]$. The identity permutation (1) fixes every $x \in X$, and so $\text{Fix}((1)) = 3^6$ (there are 3 colors). Now τ fixes a 6-tuple x if it is a "palindrome," that is, if the colors in x read the same forward as backward. For example,

$$x = (r, r, w, w, r, r)$$

is fixed by τ. Conversely, if

$$x = (c_1, c_2, c_3, c_4, c_5, c_6)$$

is fixed by $\tau = (1\ 6)(2\ 5)(3\ 4)$, then $c_1 = c_6$, $c_2 = c_5$, and $c_3 = c_4$; that is, x is a palindrome. It follows that $\text{Fix}(\tau) = 3^3$, for there are 3 choices for each of c_1, c_2, and c_3. The number of flags is thus

$$N = \tfrac{1}{2}(3^6 + 3^3) = 378.$$

Let us make the notion of coloring more precise.

Definition. If a group G acts on $X = \{1, \ldots, n\}$, and if $\mathcal{C}$ is a set of q *colors*, then G acts on the set $\mathcal{C}^n$ of all n-tuples of colors by

$$\tau(c_1, \ldots, c_n) = (c_{\tau 1}, \ldots, c_{\tau n}) \quad \text{for all } \tau \in G.$$

An orbit of $(c_1, \ldots, c_n) \in \mathcal{C}^n$ is called a ***(q, G)-coloring*** of X.

1	2	3	4
5	6	7	8
9	10	11	12
13	14	15	16

13	9	5	1
14	10	6	2
15	11	7	3
16	12	8	4

Figure 2.10

Example 2.114.
Color each square in a 4×4 grid red or black (adjacent squares may have the same color; indeed, one possibility is that all the squares have the same color).

If X consists of the 16 squares in the grid and if $\mathcal{C}$ consists of the two colors red and black, then the cyclic group $G = \langle R \rangle$ of order 4 acts on X, where R is clockwise rotation by 90°; Figure 2.10 shows how R acts: The right square is R's action on the left square. In cycle notation,

$$\begin{aligned} R &= (1,\ 4,\ 16,\ 13)(2,\ 8,\ 15,\ 9)(3,\ 12,\ 14,\ 5)(6,\ 7,\ 11,\ 10), \\ R^2 &= (1,\ 16)(4,\ 13)(2,\ 15)(8,\ 9)(3,\ 14)(12,\ 5)(6,\ 11)(7,\ 10), \\ R^3 &= (1,\ 13,\ 16,\ 4)(2,\ 9,\ 15,\ 8)(3,\ 5,\ 14,\ 12)(6,\ 10,\ 11,\ 7). \end{aligned}$$

A red-and-black chessboard does not change when it is rotated; it is merely viewed from a different position. Thus, we may regard a chessboard as a 2-coloring of X; the orbit of a 16-tuple corresponds to the four ways of viewing the board.

By Burnside's lemma, the number of chessboards is

$$\tfrac{1}{4}\Big[\text{Fix}((1)) + \text{Fix}(R) + \text{Fix}(R^2) + \text{Fix}(R^3)\Big].$$

Now $\text{Fix}((1)) = 2^{16}$, for every 16-tuple is fixed by the identity. To compute $\text{Fix}(R)$, note that squares 1, 4, 16, 13 must all have the same color in a 16-tuple fixed by R. Similarly, squares 2, 8, 15, 9 must have the same color, squares 3, 12, 14, 5 must have the same color, and squares 6, 7, 11, 10 must have the same color. We conclude that $\text{Fix}(R) = 2^4$; note that the exponent 4 is the number of cycles in the complete factorization of R. A similar analysis shows that $\text{Fix}(R^2) = 2^8$, for the complete factorization of R^2 has 8 cycles, and $\text{Fix}(R^3) = 2^4$, because the cycle structure of R^3 is the same as that of R. Therefore, the number N of chessboards is

$$N = \tfrac{1}{4}\Big[2^{16} + 2^4 + 2^8 + 2^4\Big] = 16{,}456. \quad ◀$$

We now show, as in Example 2.114, that the cycle structure of a permutation τ allows one to calculate $\text{Fix}(\tau)$.

Lemma 2.115. *Let $\mathcal{C}$ be a set of q colors, and let G be a subgroup of S_n. If $\tau \in G$, then*

$$\text{Fix}(\tau) = q^{t(\tau)},$$

where $t(\tau)$ is the number of cycles in the complete factorization of τ.

Proof. Since $\tau(c_1, \ldots, c_n) = (c_{\tau 1}, \ldots, c_{\tau n}) = (c_1, \ldots, c_n)$, we see that $c_{\tau i} = c_i$ for all i, and so τi has the same color as i. It follows, for all k, that $\tau^k i$ has the same color as i, that is, all points in the orbit of i acted on by $\langle \tau \rangle$ have the same color. If the complete factorization of τ is $\tau = \beta_1 \cdots \beta_{t(\tau)}$, and if i occurs in β_j, then Example 2.96 shows that the orbit containing i is the set of symbols occurring in β_j. Thus, for an n-tuple to be fixed by τ, all the symbols involved in each of the $t(\tau)$ cycles must have the same color; as there are q colors, there are thus $q^{t(\tau)} n$-tuples fixed by τ. •

Corollary 2.116. *Let G act on a finite set X. If N is the number of (q, G)-colorings of X, then*

$$N = \frac{1}{|G|} \sum_{\tau \in G} q^{t(\tau)},$$

where $t(\tau)$ is the number of cycles in the complete factorization of τ.

There is a generalization of this technique, due to G. Pólya (see Biggs, *Discrete Mathematics*), giving a formula, for example, that counts the number of red, white, blue, and green flags having 20 stripes exactly 7 of which are red and 5 of which are blue.

EXERCISES

2.78 If a and b are elements in a group G, prove that ab and ba have the same order.
Hint. Use a conjugation.

2.79 Prove that if G is a finite group of odd order, then no $x \in G$, other than $x = 1$, is conjugate to its inverse.
Hint. If x is conjugate to x^{-1}, how many elements are in x^G?

2.80 Prove that no pair of the following groups of order 8,

$$\mathbb{I}_8; \ \ \mathbb{I}_4 \times \mathbb{I}_2; \ \ \mathbb{I}_2 \times \mathbb{I}_2 \times \mathbb{I}_2; \ \ D_8; \ \ \mathbf{Q},$$

are isomorphic.

2.81 Prove that if p is a prime and G is a finite group in which every element has order a power of p, then G is a p-group. (A possibly infinite group G is called a ***p-group*** if every element in G has order a power of p.)
Hint. Use Cauchy's theorem.

2.82 Define the ***centralizer*** $C_G(H)$ of a subgroup $H \leq G$ to be

$$C_G(H) = \{x \in G : xh = hx \text{ for all } h \in H\}.$$

(i) For every subgroup $H \leq G$, prove that $C_G(H) \lhd N_G(H)$.

(ii) For every subgroup $H \leq G$, prove that $N_G(H)/C_G(H)$ is isomorphic to a subgroup of $\text{Aut}(H)$.

Hint. Generalize the homomorphism Γ in Exercise 2.64 on page 82.

2.83 Show that S_4 has a subgroup isomorphic to D_8.

2.84 Prove that $S_4/\mathbf{V} \cong S_3$.
Hint. Use Proposition 2.90.

2.85 (i) Prove that $A_4 \ncong D_{12}$.

Hint. Recall that A_4 has no element of order 6.

(ii) Prove that $D_{12} \cong S_3 \times \mathbb{I}_2$.

Hint. Each element $x \in D_{12}$ has a unique factorization of the form $x = b^i a$, where $b^6 = 1$ and $a^2 = 1$.

2.86 (i) If G is a group, then a normal subgroup $H \lhd G$ is called a ***maximal normal subgroup*** if there is no normal subgroup K of G with $H < K < G$. Prove that a normal subgroup H is a maximal normal subgroup of G if and only if G/H is a simple group.

(ii) Prove that every finite abelian group G has a subgroup of prime index.

Hint. Use Proposition 2.107.

(iii) Prove that A_6 has no subgroup of prime index.

2.87 Prove that $H \lhd N_G(H)$ and that $N_G(H)$ is the largest subgroup of G containing H as a normal subgroup.

2.88 Find $N_G(H)$ if $G = S_4$ and $H = \langle(1\ 2\ 3)\rangle$.

2.89 (i) If H is a subgroup of G and if $x \in H$, prove that

$$C_H(x) = H \cap C_G(x).$$

(ii) If H is a subgroup of index 2 in a finite group G and if $x \in H$, prove that $|x^H| = |x^G|$ or $|x^H| = \frac{1}{2}|x^G|$, where x^H is the conjugacy class of x in H.

Hint. Use the second isomorphism theorem.

(iii) Prove that there are two conjugacy classes of 5-cycles in A_5, each of which has 12 elements.

Hint. If $\alpha = (1\ 2\ 3\ 4\ 5)$, then $\left|C_{S_5}(\alpha)\right| = 5$ because $24 = \dfrac{120}{|C_{S_5}(\alpha)|}$; hence $C_{S_5}(\alpha) = \langle\alpha\rangle$. What is $C_{A_5}(\alpha)$?

(iv) Prove that the conjugacy classes in A_5 have sizes 1, 12, 12, 15, and 20.

2.90 (i) Prove that every normal subgroup H of a group G is a union of conjugacy classes of G, one of which is $\{1\}$.

(ii) Use part (i) and Exercise 2.89 to give a second proof of the simplicity of A_5.

2.91 (i) For all $n \geq 5$, prove that all 3- cycles are conjugate in A_n.

Hint. Show that $(1\ 2\ 3)$ and $(i\ j\ k)$ are conjugate, in two steps: First, if they are not disjoint (so the permutations move at most 5 letters); then, if they are disjoint.

(ii) Prove that if a normal subgroup $H \lhd A_n$ contains a 3-cycle, where $n \geq 5$, then $H = A_n$. (*Remark*. We have proved this in Lemma 2.109 when $n = 5$.)

2.92 Prove that the only normal subgroups of S_4 are $\{(1)\}$, $\mathbf{V}$, A_4, and S_4.
Hint. Use Theorem 2.9, checking the various cycle structures one at a time.

2.93 Prove that A_5 is a group of order 60 that has no subgroup of order 30.
Hint. Use Proposition 2.62(ii).

2.94 (i) Prove, for all $n \geq 5$, that the only normal subgroups of S_n are $\{(1)\}$, A_n, and S_n.

(ii) Prove that if $n \geq 3$, then A_n is the only subgroup of S_n of order $\frac{1}{2}n!$.
Hint. If H is a second such subgroup, then H is normal in S_n and hence $H \cap A_n$ is normal in A_n.

(iii) Prove that S_5 has no subgroup of order 30.
Hint. Use the representation on the cosets of a supposed subgroup of order 30, as well as the simplicity of A_5.

(iv) Prove that S_5 contains no subgroup of order 40.

2.95 Let G be a subgroup of S_n.

(i) If $G \cap A_n = \{1\}$, prove that $|G| \leq 2$.

(ii) If G is a simple group with more than 2 elements, prove that $G \leq A_n$.

2.96 (i) If $n \geq 5$, prove that S_n has no subgroup of index r, where $2 < r < n$.

(ii) Prove that if $n \geq 5$, then A_n has no subgroup of index r, where $2 \leq r < n$.

2.97 (i) Prove that if a simple group G has a subgroup of index $n > 1$, then G is isomorphic to a subgroup of S_n.
Hint. Kernels are normal subgroups.

(ii) Prove that an infinite simple group (such do exist) has no subgroups of finite index $n > 1$.
Hint. Use part (i).

2.98 Let G be a group with $|G| = mp$, where p is a prime and $1 < m < p$. Prove that G is not simple.
Hint. Show that G has a subgroup H of order p, and use the representation of G on the cosets of H.

Remark. Of all the numbers smaller than 60, we can now show that all but 11 are not orders of nonabelian simple groups (namely, 12, 18, 24, 30, 36, 40, 45, 48, 50, 54, 56). Theorem 2.103 eliminates all prime powers (for the center is always a normal subgroup), and this exercise eliminates all numbers of the form mp, where p is a prime and $m < p$. (We can complete the proof that there are no nonabelian simple groups of order less than 60 using Sylow's theorem; see Proposition 5.41.) ◄

2.99 (i) Let a group G act on a set X, and suppose that $x, y \in X$ lie in the same orbit: $y = gx$ for some $g \in G$. Prove that $G_y = gG_xg^{-1}$.

(ii) Let G be a finite group acting on a set X; prove that if $x, y \in X$ lie in the same orbit, then $|G_x| = |G_y|$.

2.100 How many flags are there with n stripes each of which can be colored any one of q given colors?
Hint. The parity of n is relevant.

2.101 Let X be the squares in an $n \times n$ grid, and let ρ be a rotation by $90°$. Define a ***chessboard*** to be a (q, G)-coloring, where the cyclic group $G = \langle \rho \rangle$ of order 4 is acting. Show that the number of chessboards is

$$\tfrac{1}{4}\left(q^{n^2} + q^{\lfloor (n^2+1)/2 \rfloor} + 2q^{\lfloor (n^2+3)/4 \rfloor}\right),$$

where $\lfloor x \rfloor$ is the greatest integer in the number x.

2.102 Let X be a disk divided into n congruent circular sectors, and let ρ be a rotation by $(360/n)°$. Define a ***roulette wheel*** to be a (q, G)-coloring, where the cyclic group $G = \langle \rho \rangle$ of order n is acting. Prove that if $n = 6$, then there are $\frac{1}{6}(2q + 2q^2 + q^3 + q^6)$ roulette wheels having 6 sectors.

The formula for the number of roulette wheels with n sectors is

$$\tfrac{1}{n} \sum_{d \mid n} \phi(n/d) q^d,$$

where ϕ is the Euler ϕ-function.

2.103 Let X be the vertices of a regular n-gon, and let the dihedral group $G = D_{2n}$ act (as the usual group of symmetries [see Example 2.28]). Define a ***bracelet*** to be a (q, G)-coloring of a regular n-gon, and call each of its vertices a ***bead***. (Not only can we rotate a bracelet, we can also *flip* it: that is, turn it upside down by rotating it in space about a line joining two beads.)

(i) How many bracelets are there having 5 beads, each of which can be colored any one of q available colors?

Hint. The group $G = D_{10}$ is acting. Use Example 2.28 to assign to each symmetry a permutation of the vertices, and then show that the number of bracelets is

$$\tfrac{1}{10}\left(q^5 + 4q + 5q^3\right).$$

(ii) How many bracelets are there having 6 beads, each of which can be colored any one of q available colors?

Hint. The group $G = D_{12}$ is acting. Use Example 2.28 to assign to each symmetry a permutation of the vertices, and then show that the number of bracelets is

$$\tfrac{1}{12}\left(q^6 + 2q^4 + 4q^3 + 3q^2 + 2q\right).$$

3

Commutative Rings I

3.1 Introduction

As in Chapters 1 and 2, this chapter contains some material usually found in an earlier course; proofs of such results are only sketched, but other theorems are proved in full. We begin by introducing commutative rings, the most prominent examples being $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$, and $\mathbb{C}$, as well as $\mathbb{I}_m$, polynomials, real-valued functions, and finite fields. We will also give some of the first results about vector spaces (with scalars in any field) and linear transformations. Canonical forms, which classify similar matrices, will be discussed in Chapter 9.

3.2 First Properties

We begin with the definition of commutative ring.

Definition. A ***commutative ring***[1] R is a set with two binary operations, addition and multiplication, such that

(i) R is an abelian group under addition;

(ii) (***commutativity***) $ab = ba$ for all $a, b \in R$;

(iii) (***associativity***) $a(bc) = (ab)c$ for every $a, b, c \in R$;

[1] This term was probably coined by D. Hilbert, in 1897, when he wrote *Zahlring*. One of the meanings of the word *ring*, in German as in English, is collection, as in the phrase "a ring of thieves." (It has also been suggested that Hilbert used this term because, for a ring of algebraic integers, an appropriate power of each element "cycles back" to being a linear combination of lower powers.)

(iv) there is an element $1 \in R$ with $1a = a$ for every $a \in R$;[2]

(v) (***distributivity***) $a(b + c) = ab + ac$ for every $a, b, c \in R$.

The element 1 in a ring R has several names; it is called *one*, the *unit* of R, or the *identity* in R.

Addition and multiplication in a commutative ring R are binary operations, so there are functions

$$\alpha : R \times R \to R \qquad \text{with} \qquad \alpha(r, r') = r + r' \in R$$

and

$$\mu : R \times R \to R \qquad \text{with} \qquad \mu(r, r') = rr' \in R$$

for all $r, r' \in R$. The law of substitution holds here, as it does for any operation: If $r = r'$ and $s = s'$, then $r + s = r' + s'$ and $rs = r's'$.

Example 3.1.

(i) $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$, and $\mathbb{C}$ are commutative rings with the usual addition and multiplication (the ring axioms are verified in courses in the foundations of mathematics).

(ii) $\mathbb{I}_m$, the integers mod m, is a commutative ring.

(iii) Let $\mathbb{Z}[i]$ be the set of all complex numbers of the form $a + bi$, where $a, b \in \mathbb{Z}$ and $i^2 = -1$. It is a boring exercise to check that $\mathbb{Z}[i]$ is, in fact, a commutative ring (this exercise will be significantly shortened, in Exercise 3.8 on page 124, once the notion of *subring* has been introduced). $\mathbb{Z}[i]$ is called the ring of ***Gaussian integers***.

(iv) Consider the set R of all real numbers x of the form

$$x = a + b\omega,$$

where $a, b \in \mathbb{Q}$ and $\omega = \sqrt[3]{2}$. It is easy to see that R is closed under ordinary addition. However, if R is closed under multiplication, then $\omega^2 \in R$, and there are rationals a and b with

$$\omega^2 = a + b\omega.$$

Multiplying both sides by ω and by b gives the equations

$$2 = a\omega + b\omega^2$$
$$b\omega^2 = ab + b^2\omega.$$

Hence, $2 - a\omega = ab + b^2\omega$, and so

$$2 - ab = (b^2 + a)\omega.$$

If $b^2 + a \neq 0$, then $\omega = \sqrt[3]{2}$ is rational; if $b^2 + a = 0$, then this coupled with $2 - ab = 0$ yields $2 = (-b)^3$. Thus, either case forces $\sqrt[3]{2}$ rational, and this contradiction shows that R is not a commutative ring. ◀

[2]Some authors do not demand that commutative rings have 1. For them, the set of all even integers is a commutative ring, but we do not recognize it as such.

Remark. There are noncommutative rings; that is, sets having an addition and a multiplication satisfying all the axioms of a commutative ring except the axiom: $ab = ba$. [Actually, the definition replaces the axiom $1a = a$ by $1a = a = a1$, and it replaces the distributive law by two distributive laws, one on either side: $a(b + c) = ab + ac$ and $(b + c)a = ba + ca$.] For example, it is easy to see that the set of all $n \times n$ real matrices, equipped with the usual addition and multiplication, satisfies all the new ring axioms. We shall study noncommutative rings in Chapter 8. ◄

Here are some elementary results.

Proposition 3.2. *Let R be a commutative ring.*

(i) $0 \cdot a = 0$ *for every* $a \in R$.

(ii) *If* $1 = 0$, *then* R *consists of the single element* 0. *In this case,* R *is called the* ***zero ring***.[3]

(iii) *If* $-a$ *is the additive inverse of* a, *then* $(-1)(-a) = a$.

(iv) $(-1)a = -a$ *for every* $a \in R$.

(v) *If* $n \in \mathbb{N}$ *and* $n1 = 0$, *then* $na = 0$ *for all* $a \in R$.

(vi) *The binomial theorem holds*: *If* $a, b \in R$, *then*

$$(a + b)^n = \sum_{r=0}^{n} \binom{n}{r} a^r b^{n-r}.$$

Sketch of Proof. (i) $0 \cdot a = (0 + 0) \cdot a = 0 \cdot a + 0 \cdot a$.

(ii) $a = 1 \cdot a = 0 \cdot a = 0$.

(iii) $0 = (-1 + 1)(-a) = (-1)(-a) + (-a)$.

(iv) Since $(-1)(-a) = a$, we have $(-1)(-1)(-a) = (-1)a$. But $(-1)(-1) = 1$.

(v) In Chapter 2, we defined the powers a^n of an element in a group, where $n \geq 0$. In an additive group, na is a more appropriate notation than a^n, and the notation na, for $n \in \mathbb{Z}$ and $a \in R$, has this meaning in R; that is, na is the sum of a with itself n times.

If $a \in R$ and $n \in \mathbb{Z}$ is positive, then $n1 = 0$ implies

$$na = n(1a) = (n1)a = 0a = 0.$$

(vi) Induction on $n \geq 0$ using the identity $\binom{n+1}{r} = \binom{n}{r-1} + \binom{n}{r}$ for $0 < r < n + 1$. (We agree that $a^0 = 1$ for all $a \in R$, even for $a = 0$.) •

A *subring* S of a commutative ring R is a commutative ring contained in a larger commutative ring R so that S and R have the same addition, multiplication, and unit.

[3] The zero ring is not a very interesting ring, but it does arise occasionally.

Definition. A subset S of a commutative ring R is a ***subring*** of R if

(i) $1 \in S$;[4]

(ii) if $a, b \in S$, then $a - b \in S$;

(iii) if $a, b \in S$, then $ab \in S$.

Notation. In contrast to the usage $H \leq G$ for a subgroup, the tradition in ring theory is to write $S \subseteq R$ for a subring. We shall also write $S \subsetneq R$ to denote a *proper* subring; that is, $S \subseteq R$ and $S \neq R$.

Proposition 3.3. *A subring S of a commutative ring R is itself a commutative ring.*

Sketch of Proof. The first condition says that S is a subgroup of the additive group R. The other conditions are identities that hold for all elements in R, and hence hold, in particular, in S. For example, associativity $a(bc) = (ab)c$ holds for all $a, b, c \in R$, and so it holds, in particular, for all $a, b, c \in S \subseteq R$. •

Of course, one advantage of the notion of subring is that fewer ring axioms need to be checked to determine whether a subset of a commutative ring is itself a commutative ring.

Exercise 3.4 on page 124 gives a natural example of a commutative ring S contained in a commutative ring R in which both S and R have the same addition and multiplication, but whose units are distinct (and so S is *not* a subring of R).

Example 3.4.
If $n \geq 3$ is an integer, let $\zeta_n = e^{2\pi i/n}$ be a primitive nth root of unity, and define

$$\mathbb{Z}[\zeta_n] = \{z \in \mathbb{C} : z = a_0 + a_1\zeta_n + a_2\zeta_n^2 + \cdots + a_{n-1}\zeta_n^{n-1}, \text{ all } a_i \in \mathbb{Z}\}.$$

(When $n = 4$, then $\mathbb{Z}[\zeta_4]$ is the Gaussian integers $\mathbb{Z}[i]$.) It is easy to check that $\mathbb{Z}[\zeta_n]$ is a subring of $\mathbb{C}$ (to prove that $\mathbb{Z}[\zeta_n]$ is closed under multiplication, note that if $m \geq n$, then $m = qn + r$, where $0 \leq r < n$, and $\zeta_n^m = \zeta_n^r$). ◂

Definition. A ***domain*** (often called an *integral domain*) is a commutative ring R that satisfies two extra axioms: first,

$$1 \neq 0;$$

second, the ***cancellation law*** for multiplication: For all $a, b, c \in R$,

$$\text{if } ca = cb \text{ and } c \neq 0, \text{ then } a = b.$$

The familiar examples of commutative rings, $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$, and $\mathbb{C}$, are domains; the zero ring is not a domain.

[4] The even integers do *not* form a subring of $\mathbb{Z}$ because 1 is not even. Their special structure will be recognized when *ideals* are introduced.

Proposition 3.5. *A nonzero commutative ring R is a domain if and only if the product of any two nonzero elements of R is nonzero.*

Sketch of Proof. $ab = ac$ if and only if $a(b - c) = 0$. •

Proposition 3.6. *The commutative ring $\mathbb{I}_m$ is a domain if and only if m is a prime.*

Proof. If $m = ab$, where $1 < a, b < m$, then $[a] \neq [0]$ and $[b] \neq [0]$ in $\mathbb{I}_m$, yet $[a][b] = [m] = [0]$.

Conversely, if m is prime and $[a][b] = [ab] = [0]$, then $m \mid ab$, and Euclid's lemma gives $m \mid a$ or $m \mid b$. •

Example 3.7.
(i) Let $\mathcal{F}(\mathbb{R})$ be the set of all the functions $\mathbb{R} \to \mathbb{R}$ equipped with the operations of ***pointwise addition*** and ***pointwise multiplication***: Given $f, g \in \mathcal{F}(\mathbb{R})$, define functions $f + g$ and fg by

$$f + g \colon a \mapsto f(a) + g(a) \qquad \text{and} \qquad fg \colon a \mapsto f(a)g(a)$$

(notice that fg is *not* their composite).

We claim that $\mathcal{F}(\mathbb{R})$ with these operations is a commutative ring. Verification of the axioms is left to the reader with the following hint: The zero element in $\mathcal{F}(\mathbb{R})$ is the constant function z with value 0 [that is, $z(a) = 0$ for all $a \in \mathbb{R}$] and the unit is the constant function ε with $\varepsilon(a) = 1$ for all $a \in \mathbb{R}$. We now show that $\mathcal{F}(\mathbb{R})$ is not a domain.

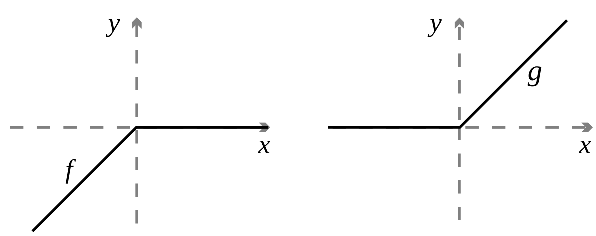


Figure 3.1

Define f and g as drawn in Figure 3.1:

$$f(a) = \begin{cases} a & \text{if } a \leq 0 \\ 0 & \text{if } a \geq 0; \end{cases} \qquad g(a) = \begin{cases} 0 & \text{if } a \leq 0 \\ a & \text{if } a \geq 0. \end{cases}$$

Clearly, neither f nor g is zero (i.e., $f \neq z$ and $g \neq z$). On the other hand, for each $a \in \mathbb{R}$, $fg \colon a \mapsto f(a)g(a) = 0$, because at least one of the factors $f(a)$ or $g(a)$ is the number zero. Therefore, $fg = z$, by Proposition 1.43, and $\mathcal{F}(\mathbb{R})$ is not a domain.

(ii) All differentiable functions $f : \mathbb{R} \to \mathbb{R}$ form a subring of $\mathcal{F}(\mathbb{R})$. The identity ε is a constant function, hence is differentiable, while the sum and product of differentiable functions are also differentiable. Hence, the differentiable functions form a commutative ring. ◄

Many theorems of ordinary arithmetic, that is, properties of the commutative ring $\mathbb{Z}$, hold in more generality. We now generalize some familiar definitions from $\mathbb{Z}$ to arbitrary commutative rings.

Definition. Let a and b be elements of a commutative ring R. Then a ***divides*** b ***in R*** (or a is a ***divisor*** of b or b is a ***multiple*** of a), denoted by $a \mid b$, if there exists an element $c \in R$ with $b = ca$.

As an extreme example, if $0 \mid a$, then $a = 0 \cdot b$ for some $b \in R$. Since $0 \cdot b = 0$, however, we must have $a = 0$. Thus, $0 \mid a$ if and only if $a = 0$.

Notice that whether $a \mid b$ depends not only on the elements a and b but on the ambient ring R as well. For example, 3 does divide 2 in $\mathbb{Q}$, for $2 = 3 \times \frac{2}{3}$, and $\frac{2}{3} \in \mathbb{Q}$; on the other hand, 3 does not divide 2 in $\mathbb{Z}$, because there is no *integer* c with $3c = 2$.

Definition. An element u in a commutative ring R is called a ***unit*** if $u \mid 1$ in R, that is, if there exists $v \in R$ with $uv = 1$; the element v is called the ***inverse*** of u and v is often denoted by u^{-1}.

Units are of interest because we can always divide by them: If $a \in R$ and u is a unit in R (so there is $v \in R$ with $uv = 1$), then

$$a = u(va)$$

is a factorization of a in R, for $va \in R$; thus, it is reasonable to define the quotient a/u as $va = u^{-1}a$.

Given elements a and b, whether $a \mid b$ depends not only on these elements but also on the ambient ring R; similarly, whether an element $u \in R$ is a unit also depends on the ambient ring R (for it is a question whether $u \mid 1$ in R). For example, the number 2 is a unit in $\mathbb{Q}$, for $\frac{1}{2}$ lies in $\mathbb{Q}$ and $2 \times \frac{1}{2} = 1$, but 2 is not a unit in $\mathbb{Z}$, because there is no *integer* v with $2v = 1$. In fact, the only units in $\mathbb{Z}$ are 1 and -1.

Proposition 3.8. *Let R be a domain, and let $a, b \in R$ be nonzero. Then $a \mid b$ and $b \mid a$ if and only if $b = ua$ for some unit $u \in R$.*

Sketch of Proof. If $b = ua$ and $a = vb$, then $b = ua = uvb$. •

There exist examples of commutative rings in which Proposition 3.8 is false, and so the hypothesis that R be a domain is needed.

What are the units in $\mathbb{I}_m$?

Proposition 3.9. *If a is an integer, then $[a]$ is a unit in $\mathbb{I}_m$ if and only if a and m are relatively prime. In fact, if $sa + tm = 1$, then $[a]^{-1} = [s]$.*

Sketch of Proof. $sa \equiv 1 \bmod m$ if and only if $sa + tm = 1$ for some integer t. •

Corollary 3.10. *If p is a prime, then every nonzero $[a]$ in $\mathbb{I}_p$ is a unit.*

Sketch of Proof. If $1 \le a < p$, then $(a, p) = 1$. •

Definition. If R is a commutative ring, then the ***group of units*** of R is

$$U(R) = \{\text{all units in } R\}.$$

It is easy to check that $U(R)$ is a multiplicative group. It follows that a unit u in R has exactly one inverse in R, for each element in a group has a unique inverse.

There is an obvious difference between $\mathbb{Q}$ and $\mathbb{Z}$: every nonzero element of $\mathbb{Q}$ is a unit.

Definition. A ***field***[5] F is a commutative ring in which $1 \ne 0$ and every nonzero element a is a unit; that is, there is $a^{-1} \in F$ with $a^{-1}a = 1$.

The first examples of fields are $\mathbb{Q}$, $\mathbb{R}$, and $\mathbb{C}$.

The definition of *field* can be restated in terms of the group of units; a commutative ring R is a field if and only if $U(R) = R^\times$, the nonzero elements of R. To say this another way, R is a field if and only if $R^\times$ is a multiplicative group [note that $U(R^\times) \ne \varnothing$ because we are assuming that $1 \ne 0$].

Proposition 3.11. *Every field F is a domain.*

Sketch of Proof. If $ab = ac$ and $a \ne 0$, then $b = a^{-1}(ab) = a^{-1}(ac) = c$. •

The converse of this proposition is false, for $\mathbb{Z}$ is a domain that is not a field.

Proposition 3.12. *The commutative ring $\mathbb{I}_m$ is a field if and only if m is prime.*

Sketch of Proof. Corollary 3.10. •

In Theorem 3.127, we shall see that there are finite fields having exactly p^n elements, whenever p is prime and $n \ge 1$; in Exercise 3.14 on page 125, we construct a field with four elements.

Every subring of a domain is itself a domain. Since fields are domains, it follows that every subring of a field is a domain. The converse of this exercise is true, and it is much more interesting: Every domain is a subring of a field.

[5]The derivation of the mathematical usage of the English term *field* (first used by E. H. Moore in 1893 in his article classifying the finite fields) as well as the German term *Körper* and the French term *corps* is probably similar to the derivation of the words *group* and *ring*: Each word denotes a "realm" or a "collection of things." The word *domain* abbreviates the usual English translation *integral domain* of the German word *Integretätsbereich*, a collection of integers.

Given four elements a, b, c, and d in a field F with $b \neq 0$ and $d \neq 0$, assume that $ab^{-1} = cd^{-1}$. Multiply both sides by bd to obtain $ad = bc$. In other words, were ab^{-1} written as a/b, then we have just shown that $a/b = c/d$ implies $ad = bc$; that is, "cross-multiplication" is valid. Conversely, if $ad = bc$ and both b and d are nonzero, then multiplication by $b^{-1}d^{-1}$ gives $ab^{-1} = cd^{-1}$, that is, $a/b = c/d$.

The proof of the next theorem is a straightforward generalization of the usual construction of the field of rational numbers $\mathbb{Q}$ from the domain of integers $\mathbb{Z}$.

Theorem 3.13. *If R is a domain, then there is a field F containing R as a subring. Moreover, F can be chosen so that, for each $f \in F$, there are $a, b \in R$ with $b \neq 0$ and $f = ab^{-1}$.*

Sketch of Proof. Let $X = \{(a, b) \in R \times R : b \neq 0\}$, and define a relation $\equiv$ on X by $(a, b) \equiv (c, d)$ if $ad = bc$. We claim that $\equiv$ is an equivalence relation. Verifications of reflexivity and symmetry are straightforward; here is the proof of transitivity. If $(a, b) \equiv (c, d)$ and $(c, d) \equiv (e, f)$, then $ad = bc$ and $cf = de$. But $ad = bc$ gives $adf = b(cf) = bde$. Canceling d, which is nonzero, gives $af = be$; that is, $(a, b) \equiv (e, f)$.

Denote the equivalence class of (a, b) by $[a, b]$, define F as the set of all equivalence classes $[a, b]$, and equip F with the following addition and multiplication (if we pretend that $[a, b]$ is the fraction a/b, then these are just the usual formulas):

$$[a, b] + [c, d] = [ad + bc, bd]$$

and

$$[a, b][c, d] = [ac, bd].$$

First, since $b \neq 0$ and $d \neq 0$, we have $bd \neq 0$, because R is a domain, and so the formulas make sense. Let us show that addition is well-defined. If $[a, b] = [a', b']$ (that is, $ab' = a'b$) and $[c, d] = [c', d']$ (that is, $cd' = c'd$), then we must show that $[ad+bc, bd] = [a'd' + b'c', b'd']$. But this is true:

$$(ad + bc)b'd' = ab'dd' + bb'cd' = a'bdd' + bb'c'd = (a'd' + b'c')bd.$$

A similar argument shows that multiplication is well-defined.

The verification that F is a commutative ring is now routine: The zero element is $[0, 1]$, the one is $[1, 1]$, and the additive inverse of $[a, b]$ is $[-a, b]$. It is easy to see that the family $R' = \{[a, 1] : a \in R\}$ is a subring of F, and we identify $a \in R$ with $[a, 1] \in R'$.

To see that F is a field, observe that if $[a, b] \neq [0, 1]$, then $a \neq 0$, and the inverse of $[a, b]$ is $[b, a]$.

Finally, if $b \neq 0$, then $[1, b] = [b, 1]^{-1}$, and so $[a, b] = [a, 1][b, 1]^{-1}$. •

Definition. The field F constructed from R in Theorem 3.13 is called the ***fraction field*** of R; we denote it by $\text{Frac}(R)$, and we denote $[a, b] \in \text{Frac}(R)$ by a/b; in particular, the elements $[a, 1]$ of R' are denoted by $a/1$ or, more simply, by a.

Notice that the fraction field of $\mathbb{Z}$ is $\mathbb{Q}$; that is, $\text{Frac}(\mathbb{Z}) = \mathbb{Q}$.

Definition. A ***subfield*** of a field K is a subring k of K that is also a field.

It is easy to see that a subset k of a field K is a subfield if and only if k is a subring that is closed under inverses; that is, if $a \in k$ and $a \neq 0$, then $a^{-1} \in k$. It is also routine to see that any intersection of subfields of K is itself a subfield of K (note that the intersection is not equal to $\{0\}$ because 1 lies in every subfield).

EXERCISES

3.1 Prove that a commutative ring R has a unique 1.

3.2 (i) Prove that subtraction in $\mathbb{Z}$ is not an associative operation.

(ii) Give an example of a commutative ring R in which subtraction is associative.

3.3 (i) If R is a domain and $a \in R$ satisfies $a^2 = a$, prove that either $a = 0$ or $a = 1$.

(ii) Show that the commutative ring $\mathcal{F}(\mathbb{R})$ in Example 3.7 contains infinitely many elements $f \neq 0, 1$ with $f^2 = f$.

3.4 (i) If X is a set, prove that the Boolean group $\mathcal{B}(X)$ in Example 2.18 with elements the subsets of X and with addition given by

$$U + V = (U - V) \cup (V - U),$$

where $U - V = \{x \in U : x \notin V\}$, is a commutative ring if one defines multiplication

$$UV = U \cap V.$$

We call $\mathcal{B}(X)$ a ***Boolean ring***.

Hint. You may use some standard facts of set theory: the distributive law: $U \cap (V \cup W) = (U \cap V) \cup (U \cap W)$; if V' denotes the complement of V, then $U - V = U \cap V'$; and the ***De Morgan law***: $(U \cap V)' = U' \cup V'$.

(ii) Prove that $\mathcal{B}(X)$ contains exactly one unit.

(iii) If Y is a proper subset of X (that is, $Y \subsetneq X$), show that the unit in $\mathcal{B}(Y)$ is distinct from the unit in $\mathcal{B}(X)$. Conclude that $\mathcal{B}(Y)$ is *not* a subring of $\mathcal{B}(X)$.

3.5 Show that $U(\mathbb{I}_m) = \{[k] \in \mathbb{I}_m : (k, m) = 1\}$.

3.6 Find all the units in the commutative ring $\mathcal{F}(\mathbb{R})$ defined in Example 3.7.

3.7 Generalize the construction of $\mathcal{F}(\mathbb{R})$ to arbitrary commutative rings R: Let $\mathcal{F}(R)$ be the set of all functions from R to R, with pointwise addition, $f + g \colon r \mapsto f(r) + g(r)$, and pointwise multiplication, $fg \colon r \mapsto f(r)g(r)$ for $r \in R$.

(i) Show that $\mathcal{F}(R)$ is a commutative ring.

(ii) Show that $\mathcal{F}(R)$ is not a domain.

(iii) Show that $\mathcal{F}(\mathbb{I}_2)$ has exactly four elements, and that $f + f = 0$ for every $f \in \mathcal{F}(\mathbb{I}_2)$.

3.8 (i) If R is a domain and S is a subring of R, then S is a domain.

(ii) Prove that $\mathbb{C}$ is a domain, and conclude that the ring of Gaussian integers is a domain.

3.9 Prove that the only subring of $\mathbb{Z}$ is $\mathbb{Z}$ itself.
Hint. Every subring R of $\mathbb{Z}$ contains 1.

3.10 (i) Prove that $R = \{a + b\sqrt{2} : a, b \in \mathbb{Z}\}$ is a domain.

(ii) Prove that $R = \{\frac{1}{2}(a + b\sqrt{2}) : a, b \in \mathbb{Z}\}$ is not a domain.

(iii) Using the fact that $\alpha = \frac{1}{2}(1 + \sqrt{-19})$ is a root of $x^2 - x + 5$, prove that $R = \{a + b\alpha : a, b \in \mathbb{Z}\}$ is a domain.

3.11 Prove that the set of all C^∞-functions is a subring of $\mathcal{F}(\mathbb{R})$. (A function $f : \mathbb{R} \to \mathbb{R}$ is a C^∞*-function* if it has an nth derivative $f^{(n)}$ for every $n \geq 1$.)
Hint. Use the Leibniz rule (see Exercise 1.6 on page 12).

3.12 (i) If R is a commutative ring, define the ***circle operation*** $a \circ b$ by

$$a \circ b = a + b - ab.$$

Prove that the circle operation is associative and that $0 \circ a = a$ for all $a \in R$.

(ii) Prove that a commutative ring R is a field if and only if $\{r \in R : r \neq 1\}$ is an abelian group under the circle operation.

Hint. If $a \neq 0$, then $a + 1 \neq 1$.

3.13 Find the inverses of the nonzero elements of $\mathbb{I}_{11}$.

3.14 (***R. A. Dean***) Define $\mathbb{F}_4$ to be all 2×2 matrices of the form

$$\begin{bmatrix} a & b \\ b & a + b \end{bmatrix},$$

where $a, b \in \mathbb{I}_2$.

(i) Prove that $\mathbb{F}_4$ is a commutative ring under the usual matrix operations of addition and multiplication.

(ii) Prove that $\mathbb{F}_4$ is a field with exactly four elements.

3.15 Prove that every domain R with a finite number of elements must be a field. (Using Proposition 3.6, this gives a new proof of sufficiency in Proposition 3.12.)
Hint. If $R^\times$ denotes the set of nonzero elements of R, prove that multiplication by r is an injection $R^\times \to R^\times$, where $r \in R^\times$.

3.16 Show that $F = \{a + b\sqrt{2} : a, b \in \mathbb{Q}\}$ is a field.

3.17 (i) Show that $F = \{a + bi : a, b \in \mathbb{Q}\}$ is a field.

(ii) Show that F is the fraction field of the Gaussian integers.

3.18 If R is a commutative ring, define a relation $\equiv$ on R by $a \equiv b$ if there is a unit $u \in R$ with $b = ua$. Prove that if $a \equiv b$, then $(a) = (b)$, where $(a) = \{ra : r \in R\}$. Conversely, prove that if R is a domain, then $(a) = (b)$ implies $a \equiv b$.

3.19 (i) For any field k, prove that ***stochastic group*** $\Sigma(2, k)$, the set of all nonsingular 2×2 matrices with entries in k whose column sums are 1, is a group under matrix multiplication.

(ii) Define the ***affine group*** $\text{Aff}(1, k)$ to be the set of all $f : k \to k$ of the form $f(x) = ax + b$, where $a, b \in k$ and $a \neq 0$. Prove that $\Sigma(2, k) \cong \text{Aff}(1, k)$. (See Exercise 2.46 on page 80.)

(iii) If k is a finite field with q elements, prove that $|\Sigma(2, k)| = q(q - 1)$.

(iv) Prove that $\Sigma(2, \mathbb{I}_3) \cong S_3$.

3.3 POLYNOMIALS

Even though the reader is familiar with polynomials, we now introduce them carefully. The key observation is that one should pay attention to where the coefficients of polynomials live.

Definition. If R is a commutative ring, then a ***sequence*** σ in R is

$$\sigma = (s_0, s_1, s_2, \ldots, s_i, \ \ldots);$$

the entries $s_i \in R$, for all $i \geq 0$, are called the ***coefficients*** of σ.

To determine when two sequences are equal, let us recognize that a sequence σ is really a function $\sigma : \mathbb{N} \to R$, where $\mathbb{N}$ is the set of natural numbers, with $\sigma(i) = s_i$ for all $i \geq 0$. Thus, if $\tau = (t_0, t_1, t_2, \ldots, t_i, \ldots)$ is a sequence, then $\sigma = \tau$ if and only if $\sigma(i) = \tau(i)$ for all $i \geq 0$; that is, $\sigma = \tau$ if and only if $s_i = t_i$ for all $i \geq 0$.

Definition. A sequence $\sigma = (s_0, s_1, \ldots, s_i, \ldots)$ in a commutative ring R is called a ***polynomial*** if there is some integer $m \geq 0$ with $s_i = 0$ for all $i > m$; that is,

$$\sigma = (s_0, s_1, \ldots, s_m, 0, 0, \ldots).$$

A polynomial has only finitely many nonzero coefficients. The ***zero polynomial***, denoted by $\sigma = 0$, is the sequence $\sigma = (0, 0, 0, \ldots)$.

Definition. If $\sigma = (s_0, s_1, \ldots, s_n, 0, 0, \ldots) \neq 0$ is a polynomial, then there is $s_n \neq 0$ with $s_i = 0$ for all $i > n$. We call s_n the ***leading coefficient*** of σ, we call n the ***degree*** of σ, and we denote the degree n by $\deg(\sigma)$.

The zero polynomial 0 does not have a degree because it has no nonzero coefficients. Some authors define $\deg(0) = -\infty$, and this is sometimes convenient, for $-\infty < n$ for every integer n. On the other hand, we choose not to assign a degree to 0 because it is often a genuinely different case that must be dealt with separately.

Notation. If R is a commutative ring, then the set of all polynomials with coefficients in R is denoted by $R[x]$.

Proposition 3.14. *If R is a commutative ring, then $R[x]$ is a commutative ring that contains R as a subring.*

Sketch of Proof. Define addition and multiplication of polynomials as follows: If $\sigma = (s_0, s_1, \ldots)$ and $\tau = (t_0, t_1, \ldots)$, then

$$\sigma + \tau = (s_0 + t_0, s_1 + t_1, \ldots, s_n + t_n, \ldots)$$

and

$$\sigma\tau = (c_0, c_1, c_2, \ldots),$$

where $c_k = \sum_{i+j=k} s_i t_j = \sum_{i=0}^{k} s_i t_{k-i}$. Verification of the axioms in the definition of commutative ring is routine. The subset $\{(r, 0, 0, \ldots) : r \in R\}$ is a subring of $R[x]$ that we identify with R. •

Lemma 3.15. *Let R be a commutative ring and let $\sigma, \tau \in R[x]$ be nonzero polynomials.*

(i) *Either $\sigma\tau = 0$ or $\deg(\sigma\tau) \leq \deg(\sigma) + \deg(\tau)$.*

(ii) *If R is a domain, then $\sigma\tau \neq 0$ and*

$$\deg(\sigma\tau) = \deg(\sigma) + \deg(\tau).$$

(iii) *If R is a domain, then $R[x]$ is a domain.*

Sketch of Proof. Let $\sigma = (s_0, s_1, \ldots)$ and $\tau = (t_0, t_1, \ldots)$ have degrees m and n, respectively.
(i) If $k > m + n$, then each term in $\sum_i s_i t_{k-i}$ is 0 (for either $s_i = 0$ or $t_{k-i} = 0$).

(ii) Each term in $\sum_i s_i t_{m+n-i}$ is 0, with the possible exception of $s_m t_n$. Since R is a domain, $s_m \neq 0$ and $t_n \neq 0$ imply $s_m t_n \neq 0$.

(iii) This follows from part (ii) because the product of two nonzero polynomials is now nonzero. •

Definition. If R is a commutative ring, then $R[x]$ is called the ***ring of polynomials over*** R.

Here is the link between this discussion and the usual notation.

Definition. Define the element $x \in R[x]$ by

$$x = (0, 1, 0, 0, \ldots).$$

Lemma 3.16.

(i) *If $\sigma = (s_0, s_1, \ldots)$, then*

$$x\sigma = (0, s_0, s_1, \ldots);$$

that is, multiplying by x shifts each coefficient one step to the right.

(ii) *If $n \geq 1$, then x^n is the polynomial having 0 everywhere except for 1 in the nth coordinate.*

(iii) *If $r \in R$, then*

$$(r, 0, 0, \ldots)(s_0, s_1, \ldots, s_j, \ldots) = (rs_0, rs_1, \ldots, rs_j, \ldots).$$

Sketch of Proof. Each is a routine computation using the definition of polynomial multiplication. •

If we identify $(r, 0, 0, \ldots)$ with r, then Lemma 3.16(iii) reads

$$r(s_0, s_1, \ldots, s_i, \ldots) = (rs_0, rs_1, \ldots, rs_i, \ldots).$$

We can now recapture the usual notation.

Proposition 3.17. *If $\sigma = (s_0, s_1, \ldots, s_n, 0, 0, \ldots)$, then*

$$\sigma = s_0 + s_1 x + s_2 x^2 + \cdots + s_n x^n,$$

where each element $s \in R$ is identified with the polynomial $(s, 0, 0, \ldots)$.

Proof.

$$\begin{aligned}\sigma &= (s_0, s_1, \ldots, s_n, 0, 0, \ldots)\\ &= (s_0, 0, 0, \ldots) + (0, s_1, 0, \ldots) + \cdots + (0, 0, \ldots, s_n, 0, \ldots)\\ &= s_0(1, 0, 0, \ldots) + s_1(0, 1, 0, \ldots) + \cdots + s_n(0, 0, \ldots, 1, 0, \ldots)\\ &= s_0 + s_1 x + s_2 x^2 + \cdots + s_n x^n. \quad \bullet\end{aligned}$$

We shall use this familiar (and standard) notation from now on. As is customary, we shall write

$$f(x) = s_0 + s_1 x + s_2 x^2 + \cdots + s_n x^n$$

instead of $\sigma = (s_0, s_1, \ldots, s_n, 0, 0, \ldots)$.

Here is some standard vocabulary associated with polynomials. If $f(x) = s_0 + s_1 x + s_2 x^2 + \cdots + s_n x^n$, where $s_n \neq 0$, then s_0 is called its ***constant term*** and, as we have already said, s_n is called its ***leading coefficient***. If its leading coefficient $s_n = 1$, then $f(x)$ is called ***monic***. Every polynomial other than the zero polynomial 0 (having all coefficients 0) has a degree. A ***constant polynomial*** is either the zero polynomial or a polynomial of degree 0. Polynomials of degree 1, namely, $a + bx$ with $b \neq 0$, are called ***linear***, polynomials of degree 2 are ***quadratic***,[6] degree 3's are ***cubic***, then ***quartics***, ***quintics***, and so on.

Corollary 3.18. *Polynomials $f(x) = s_0 + s_1 x + s_2 x^2 + \cdots + s_n x^n$ and $g(x) = t_0 + t_1 x + t_2 x^2 + \cdots + t_m x^m$ of degrees n and m, respectively, are equal if and only if $n = m$ and $s_i = t_i$ for all i.*

Proof. This is merely a restatement of the definition of equality of sequences, rephrased in the usual notation for polynomials. $\bullet$

[6]Quadratic polynomials are so called because the particular quadratic x^2 gives the area of a square (*quadratic* comes from the Latin word meaning "four," which is to remind us of the four-sided figure); similarly, cubic polynomials are so called because x^3 gives the volume of a cube. Linear polynomials are so called because the graph of a linear polynomial in $\mathbb{R}[x]$ is a line.

We can now describe the usual role of x in $f(x)$ as a variable. If R is a commutative ring, each polynomial $f(x) = s_0+s_1x+s_2x^2+\cdots+s_nx^n \in R[x]$ defines a ***polynomial function*** $f : R \to R$ by evaluation: If $a \in R$, define $f(a) = s_0 + s_1a + s_2a^2 + \cdots + s_na^n \in R$. The reader should realize that polynomials and polynomial functions are distinct objects. For example, if R is a finite ring (e.g., $R = \mathbb{I}_m$), then there are only finitely many functions from R to itself, and so there are only finitely many polynomial functions. On the other hand, there are infinitely many polynomials: for example, all the powers $1, x, x^2, \ldots, x^n, \ldots$ are distinct, by Corollary 3.18.

Definition. Let k be a field. The fraction field of $k[x]$, denoted by $k(x)$, is called the ***field of rational functions*** over k.

Proposition 3.19. *If k is a field, then the elements of $k(x)$ have the form $f(x)/g(x)$, where $f(x), g(x) \in k[x]$ and $g(x) \neq 0$.*

Sketch of Proof. Theorem 3.13. •

Proposition 3.20. *If p is a prime, then the field of rational functions $\mathbb{I}_p(x)$ is an infinite field containing $\mathbb{I}_p$ as a subfield.*[7]

Proof. By Lemma 3.15(iii), $\mathbb{I}_p[x]$ is an infinite domain, for the powers x^n, for $n \in \mathbb{N}$, are distinct. Thus, its fraction field, $\mathbb{I}_p(x)$, is an infinite field containing $\mathbb{I}_p[x]$ as a subring. But $\mathbb{I}_p[x]$ contains $\mathbb{I}_p$ as a subring, by Proposition 3.14. •

In spite of the difference between polynomials and polynomial functions (we shall see, in Corollary 3.28, that these objects coincide when the coefficient ring R is an infinite field), $R[x]$ is often called the ring of all *polynomials over R in one variable*. If we write $A = R[x]$, then the polynomial ring $A[y]$ is called the ring of all *polynomials over R in two variables x and y*, and it is denoted by $R[x, y]$. For example, the quadratic polynomial $ax^2 + bxy + cy^2 + dx + ey + f$ can be written $cy^2 + (bx + e)y + (ax^2 + dx + f)$, a polynomial in y with coefficients in $R[x]$. By induction, we can form the commutative ring $R[x_1, x_2, \ldots, x_n]$ of all ***polynomials in n variables*** with coefficients in R. Lemma 3.15(iii) can now be generalized, by induction on n, to say that if R is a domain, then so is $R[x_1, x_2, \ldots, x_n]$. Moreover, when k is a field, we can describe $\text{Frac}(k[x_1, x_2, \ldots, x_n])$ as all ***rational functions in n variables***; its elements have the form $f(x_1, x_2, \ldots, x_n)/g(x_1, x_2, \ldots, x_n)$, where f and g lie in $k[x_1, x_2, \ldots, x_n]$.

EXERCISES

3.20 Show that if R is a commutative ring, then $R[x]$ is never a field.
Hint. If x^{-1} exists, what is its degree?

[7] In the future, we will denote $\mathbb{I}_p$ by $\mathbb{F}_p$ when it is to be viewed as a field.

3.21 (i) If R is a domain, show that if a polynomial in $R[x]$ is a unit, then it is a nonzero constant (the converse is true if R is a field).

Hint. Compute degrees.

(ii) Show that $(2x+1)^2 = 1$ in $\mathbb{I}_4[x]$. Conclude that the hypothesis in part (i) that R be a domain is necessary.

3.22 Show that the polynomial function defined by $f(x) = x^p - x \in \mathbb{I}_p[x]$ is identically zero.

Hint. Use Fermat's theorem.

3.23 If R is a commutative ring and $f(x) = \sum_{i=0}^n s_i x^i \in R[x]$ has degree $n \geq 1$, define its ***derivative*** $f'(x) \in R[x]$ by

$$f'(x) = s_1 + 2s_2x + 3s_3x^2 + \cdots + ns_nx^{n-1};$$

if $f(x)$ is a constant polynomial, define its derivative to be the zero polynomial. Prove that the usual rules of calculus hold:

$$(f+g)' = f' + g';$$
$$(rf)' = r(f') \quad \text{if } r \in R;$$
$$(fg)' = fg' + f'g;$$
$$(f^n)' = nf^{n-1}f' \quad \text{for all } n \geq 1.$$

3.24 Let R be a commutative ring and let $f(x) \in R[x]$.

(i) Prove that if $(x-a)^2 \mid f(x)$, then $x - a \mid f'(x)$ in $R[x]$.

(ii) Prove that if $x - a \mid f(x)$ and $x - a \mid f'(x)$, then $(x-a)^2 \mid f(x)$.

3.25 (i) If $f(x) = ax^{2p} + bx^p + c \in \mathbb{I}_p[x]$, prove that $f'(x) = 0$.

(ii) Prove that a polynomial $f(x) \in \mathbb{I}_p[x]$ has $f'(x) = 0$ if and only if there is a polynomial $g(x) = \sum a_n x^n$ with $f(x) = g(x^p)$; that is, $f(x) = \sum a_n x^{np} \in \mathbb{I}_p[x^p]$.

3.26 If R is a commutative ring, define $R[[x]]$ to be the set of all sequences $(s_0, s_1, \ldots)$ with $s_i \in R$ for all i (we do not assume here that $s_i = 0$ for large i).

(i) Show that the formulas defining addition and multiplication on $R[x]$ make sense for $R[[x]]$, and prove that $R[[x]]$ is a commutative ring under these operations ($R[[x]]$ is called the ***ring of formal power series over*** R.)

(ii) Prove that $R[x]$ is a subring of $R[[x]]$.

(iii) Prove that if R is a domain, then $R[[x]]$ is a domain.

Hint. If $\sigma = (s_0, s_1, \ldots) \in R[[x]]$ is nonzero, define the ***order*** of σ, denoted by $\text{ord}(\sigma)$, to be the smallest $n \geq 0$ for which $s_n \neq 0$. If R is a domain and $\sigma, \tau \in R[[x]]$ are nonzero, prove that $\text{ord}(\sigma\tau) = \text{ord}(\sigma) + \text{ord}(\tau) \neq 0$, and hence $\sigma\tau \neq 0$.

3.27 (i) Denote a formal power series $\sigma = (s_0, s_1, s_2, \ldots, s_n, \ldots)$ by

$$\sigma = s_0 + s_1x + s_2x^2 + \cdots.$$

Prove that if $\sigma = 1 + x + x^2 + \cdots$, then $\sigma = 1/(1-x)$ in $R[[x]]$; that is, $(1-x)\sigma = 1$.

(ii) Prove that if k is a field, then a formal power series $\sigma \in k[[x]]$ is a unit if and only if its constant term is nonzero; that is, $\text{ord}(\sigma) = 0$.

(iii) Prove that if $\sigma \in k[[x]]$ and $\text{ord}(\sigma) = n$, then

$$\sigma = x^n u,$$

where u is a unit in $k[[x]]$.

3.4 GREATEST COMMON DIVISORS

We are now going to see that, when k is a field, virtually all the familiar theorems proved for $\mathbb{Z}$ have polynomial analogs in $k[x]$; moreover, we shall see that the familiar proofs can be translated into proofs here.

The division algorithm for polynomials with coefficients in a field says that long division is possible.

Theorem 3.21 (Division Algorithm). *Assume that k is a field and that $f(x), g(x) \in k[x]$ with $f(x) \neq 0$. Then there are unique polynomials $q(x), r(x) \in k[x]$ with*

$$g(x) = q(x)f(x) + r(x)$$

and either $r(x) = 0$ or $\deg(r) < \deg(f)$.

Proof. We first prove the existence of such q and r. If $f \mid g$, then $g = qf$ for some q; define the remainder $r = 0$, and we are done. If $f \nmid g$, then consider all (necessarily nonzero) polynomials of the form $g - qf$ as q varies over $k[x]$. The least integer axiom provides a polynomial $r = g - qf$ having least degree among all such polynomials. Since $g = qf + r$, it suffices to show that $\deg(r) < \deg(f)$. Write $f(x) = s_n x^n + \cdots + s_1 x + s_0$ and $r(x) = t_m x^m + \cdots + t_1 x + t_0$. Now $s_n \neq 0$ implies that s_n is a unit, because k is a field, and so s_n^{-1} exists in k. If $\deg(r) \geq \deg(f)$, define

$$h(x) = r(x) - t_m s_n^{-1} x^{m-n} f(x);$$

that is, if $\text{LT}(f) = s_n x^n$, where LT abbreviates ***leading term***, then

$$h = r - \frac{\text{LT}(r)}{\text{LT}(f)} f;$$

note that $h = 0$ or $\deg(h) < \deg(r)$. If $h = 0$, then $r = [\text{LT}(r)/\text{LT}(f)]f$ and

$$\begin{aligned} g &= qf + r \\ &= qf + \frac{\text{LT}(r)}{\text{LT}(f)} f \\ &= \left[q + \frac{\text{LT}(r)}{\text{LT}(f)}\right] f, \end{aligned}$$

contradicting $f \nmid g$. If $h \neq 0$, then $\deg(h) < \deg(r)$ and

$$g - qf = r = h + \frac{\text{LT}(r)}{\text{LT}(f)} f.$$

Thus, $g - \big[q + \text{LT}(r)/\text{LT}(f)\big] f = h$, contradicting r being a polynomial of least degree having this form. Therefore, $\deg(r) < \deg(f)$.

To prove uniqueness of $q(x)$ and $r(x)$, assume that $g = q'f + r'$, where $\deg(r') < \deg(f)$. Then

$$(q - q')f = r' - r.$$

If $r' \neq r$, then each side has a degree. But $\deg((q-q')f) = \deg(q-q') + \deg(f) \geq \deg(f)$, while $\deg(r'-r) \leq \max\{\deg(r'), \deg(r)\} < \deg(f)$, a contradiction. Hence, $r' = r$ and $(q-q')f = 0$. As $k[x]$ is a domain and $f \neq 0$, it follows that $q - q' = 0$ and $q = q'$. •

Definition. If $f(x)$ and $g(x)$ are polynomials in $k[x]$, where k is a field, then the polynomials $q(x)$ and $r(x)$ occurring in the division algorithm are called the ***quotient*** and the ***remainder*** after dividing $g(x)$ by $f(x)$.

The hypothesis that k is a field is much too strong; long division can be carried out in $R[x]$ for every commutative ring R as long as the leading coefficient of $f(x)$ is a unit in R; in particular, long division is always possible when $f(x)$ is a monic polynomial.

Corollary 3.22. *Let R be a commutative ring, and let $f(x) \in R[x]$ be a monic polynomial. If $g(x) \in R[x]$, then there exist $q(x), r(x) \in R[x]$ with*

$$g(x) = q(x)f(x) + r(x),$$

where either $r(x) = 0$ *or* $\deg(r) < \deg(f)$.

Sketch of Proof. The proof of the division algorithm can be repeated here, once we observe that $\mathrm{LT}(r)/\mathrm{LT}(f) \in R$ because $f(x)$ is monic. •

We now turn our attention to roots of polynomials.

Definition. If $f(x) \in k[x]$, where k is a field, then a ***root*** of $f(x)$ ***in k*** is an element $a \in k$ with $f(a) = 0$.

Remark. The polynomial $f(x) = x^2 - 2$ has its coefficients in $\mathbb{Q}$, but we usually say that $\sqrt{2}$ is a root of $f(x)$ even though $\sqrt{2}$ is irrational; that is, $\sqrt{2} \notin \mathbb{Q}$. We shall see later, in Theorem 3.123, that for every polynomial $f(x) \in k[x]$, where k is any field, there is a larger field E that contains k as a subfield and that contains all the roots of $f(x)$. For example, $x^2 - 2 \in \mathbb{I}_3[x]$ has no root in $\mathbb{I}_3$, but we shall see that a version of $\sqrt{2}$ does exist in some (finite) field containing $\mathbb{I}_3$. ◂

We will use the following elementary exercise in the proof of the next lemma. If $f(x), g(x) \in R[x]$, where R is a commutative ring, write

$$a(x) = f(x) + g(x) \quad \text{and} \quad m(x) = f(x)g(x);$$

evaluating at $u \in R$ gives $a(u) = f(u) + g(u)$ and $m(u) = f(u)g(u)$.

Lemma 3.23. *Let $f(x) \in k[x]$, where k is a field, and let $u \in k$. Then there is $q(x) \in k[x]$ with*

$$f(x) = q(x)(x - u) + f(u).$$

Proof. The division algorithm gives

$$f(x) = q(x)(x - u) + r;$$

the remainder r is a constant because $x - u$ has degree 1. Now evaluate:

$$f(u) = q(u)(u - u) + r,$$

and so $r = f(u)$. •

There is a connection between roots and factoring.

Proposition 3.24. *If $f(x) \in k[x]$, where k is a field, then a is a root of $f(x)$ in k if and only if $x - a$ divides $f(x)$ in $k[x]$.*

Proof. If a is a root of $f(x)$ in k, then $f(a) = 0$ and the lemma gives $f(x) = q(x)(x-a)$. Conversely, if $f(x) = g(x)(x-a)$, then evaluating at a gives $f(a) = g(a)(a-a) = 0$. •

Theorem 3.25. *Let k be a field and let $f(x) \in k[x]$. If $f(x)$ has degree n, then $f(x)$ has at most n roots in k.*

Proof. We prove the statement by induction on $n \geq 0$. If $n = 0$, then $f(x)$ is a nonzero constant, and so the number of its roots in k is zero. Now let $n > 0$. If $f(x)$ has no roots in k, then we are done, for $0 \leq n$. Otherwise, we may assume that there is $a \in k$ with a a root of $f(x)$; hence, by Proposition 3.24,

$$f(x) = q(x)(x - a);$$

moreover, $q(x) \in k[x]$ has degree $n - 1$. If there is a root $b \in k$ with $b \neq a$, then

$$0 = f(b) = q(b)(b - a).$$

Since $b - a \neq 0$, we have $q(b) = 0$ (because k is a field, hence is a domain), so that b is a root of $q(x)$. Now $\deg(q) = n - 1$, so that the inductive hypothesis says that $q(x)$ has at most $n - 1$ roots in k. Therefore, $f(x)$ has at most n roots in k. •

Example 3.26.
Theorem 3.25 is not true for polynomials with coefficients in an arbitrary commutative ring R. For example, if $R = \mathbb{I}_8$, then the quadratic polynomial $x^2 - 1 \in \mathbb{I}_8[x]$ has 4 roots: [1], [3], [5], and [7]. ◄

Corollary 3.27. *Every nth root of unity in $\mathbb{C}$ is equal to*

$$e^{2\pi ik/n} = \cos\left(\tfrac{2\pi k}{n}\right) + i \sin\left(\tfrac{2\pi k}{n}\right),$$

where $k = 0, 1, 2, \ldots, n-1$.

Proof. We have seen, in Corollary 1.35, that each of the n different complex numbers $e^{2\pi ik/n}$ is an nth root of unity; that is, each is a root of $x^n - 1$. By Theorem 3.25, there can be no other complex roots. •

Recall that every polynomial $f(x) \in k[x]$ determines the polynomial function $k \to k$ that sends a into $f(a)$ for all $a \in k$. In Exercise 3.22 on page 130, however, we saw that a nonzero polynomial in $\mathbb{I}_p[x]$ (e.g., $x^p - x$) can determine the constant function zero. This pathology vanishes when the field k is infinite.

Corollary 3.28. *Let k be an infinite field and let $f(x)$ and $g(x)$ be polynomials in $k[x]$. If $f(x)$ and $g(x)$ determine the same polynomial function [i.e., if $f(a) = g(a)$ for all $a \in k$], then $f(x) = g(x)$.*

Proof. If $f(x) \neq g(x)$, then the polynomial $h(x) = f(x) - g(x)$ is nonzero, so that it has some degree, say, n. Now every element of k is a root of $h(x)$; since k is infinite, $h(x)$ has more than n roots, and this contradicts the theorem. •

This proof yields a more general result.

Corollary 3.29. *Let k be any field, perhaps finite. If $f(x), g(x) \in k[x]$, if $\deg(f) \leq \deg(g) \leq n$, and if $f(a) = g(a)$ for $n+1$ elements $a \in k$, then $f(x) = g(x)$.*

Sketch of Proof. If $f \neq g$, then $\deg(f - g)$ is defined and $\deg(f - g) \leq n$. •

Here is another nice application of Theorem 3.25.

Theorem 3.30. *If k is a field and G is a finite subgroup of the multiplicative group $k^\times$, then G is cyclic. In particular, if k itself is finite (e.g., $k = \mathbb{I}_p$), then $k^\times$ is cyclic.*

Proof. Let d be a divisor of $|G|$. If there are two subgroups of G of order d, say, S and T, then $|S \cup T| > d$. But each $a \in S \cup T$ satisfies $a^d = 1$, by Lagrange's theorem, and hence it is a root of $x^d - 1$. This contradicts Theorem 3.25, for this polynomial now has too many roots in k. Thus, G is cyclic, by Theorem 2.86. •

Definition. If k is a finite field, a generator of the cyclic group $k^\times$ is called a ***primitive element*** of k.

Although the multiplicative groups $\mathbb{I}_p^\times$ are cyclic, no explicit formula giving a primitive element of each of them is known. For example, finding a primitive element of $\mathbb{F}_{257}$ essentially involves checking the powers of each $[i]$, where $1 < i < 257$, until one is found for which $i^m \not\equiv 1 \bmod 257$ for all positive integers $m < 256$.

The definition of a greatest common divisor of polynomials is essentially the same as the corresponding definition for integers.

Definition. If $f(x)$ and $g(x)$ are polynomials in $k[x]$, where k is a field, then a ***common divisor*** is a polynomial $c(x) \in k[x]$ with $c(x) \mid f(x)$ and $c(x) \mid g(x)$. If $f(x)$ and $g(x)$ in $k[x]$ are not both 0, define their ***greatest common divisor***, abbreviated gcd, to be the monic common divisor having largest degree. If $f(x) = 0 = g(x)$, define their gcd $= 0$. The gcd of $f(x)$ and $g(x)$ [which is uniquely determined by $f(x)$ and $g(x)$] is often denoted by (f, g).

Theorem 3.31. *If k is a field and $f(x)$, $g(x) \in k[x]$, then their* gcd $d(x)$ *is a linear combination of $f(x)$ and $g(x)$; that is, there are $s(x), t(x) \in k[x]$ with*

$$d(x) = s(x)f(x) + t(x)g(x).$$

Sketch of Proof. This proof is very similar to the corresponding result in $\mathbb{Z}$; indeed, once we introduce principal ideal domains, we will prove this theorem and its analog in $\mathbb{Z}$ simultaneously (see Theorem 3.57). •

Corollary 3.32. *Let k be a field and let $f(x)$, $g(x) \in k[x]$. A monic common divisor $d(x)$ is the* gcd *if and only if $d(x)$ is divisible by every common divisor; that is, if $c(x)$ is a common divisor, then $c(x) \mid d(x)$.*

Moreover, $f(x)$ and $g(x)$ have a unique gcd*.*

Sketch of Proof. Analogous to the proof of Proposition 1.8. •

Every polynomial $f(x)$ is divisible by u and by $uf(x)$, where u is a unit. The analog of a prime number is a polynomial having only divisors of these trivial sorts.

Definition. An element p in a domain R is ***irreducible*** if p is neither 0 nor a unit and, in any factorization $p = uv$ in R, either u or v is a unit. Elements $a, b \in R$ are ***associates*** if there is a unit $u \in R$ with $b = ua$.

For example, a prime $p \in \mathbb{Z}$ is an irreducible element, as is $-p$. We now describe irreducible polynomials $p(x) \in k[x]$, when k is a field.

Proposition 3.33. *If k is a field, then a polynomial $p(x) \in k[x]$ is irreducible if and only if* $\deg(p) = n \geq 1$ *and there is no factorization in $k[x]$ of the form $p(x) = g(x)h(x)$ in which both factors have degree smaller than n.*

Proof. We show first that $h(x) \in k[x]$ is a unit if and only if $\deg(h) = 0$. If $h(x)u(x) = 1$, then $\deg(h) + \deg(u) = \deg(1) = 0$; since degrees are nonnegative, we have $\deg(h) = 0$. Conversely, if $\deg(h) = 0$, then $h(x)$ is a nonzero constant; that is, $h \in k$; since k is a field, h has an inverse.

If $p(x)$ is irreducible, then its only factorizations are of the form $p(x) = g(x)h(x)$, where $g(x)$ or $h(x)$ is a unit; that is, where either $\deg(g) = 0$ or $\deg(h) = 0$. Therefore, $p(x)$ has no factorization in which both factors have smaller degree.

Conversely, if $p(x)$ is not irreducible, then it has a factorization $p(x) = g(x)h(x)$ in which neither $g(x)$ nor $h(x)$ is a unit; that is, neither $g(x)$ nor $h(x)$ has degree 0. Therefore, $p(x)$ has a factorization as a product of polynomials of smaller degree. •

If k is not a field, however, then this characterization of irreducible polynomials no longer holds. For example, $2x + 2 = 2(x + 1)$ is not irreducible in $\mathbb{Z}[x]$, even though, in any factorization, one factor has degree 0 and the other degree 1 (when k is a field, the units are the nonzero constants, but this is no longer true for more general coefficients).

As the definition of divisibility depends on the ambient ring, so irreducibility of a polynomial $p(x) \in k[x]$ also depends on the commutative ring $k[x]$ and hence on the field k. For example, $p(x) = x^2 + 1$ is irreducible in $\mathbb{R}[x]$, but it factors as $(x + i)(x - i)$ in $\mathbb{C}[x]$. On the other hand, a linear polynomial $f(x)$ is always irreducible [if $f = gh$, then $1 = \deg(f) = \deg(g) + \deg(h)$, and so one of g or h must have degree 0 while the other has degree $1 = \deg(f)$].

Corollary 3.34. *Let k be a field and let $f(x) \in k[x]$ be a quadratic or cubic polynomial. Then $f(x)$ is irreducible in $k[x]$ if and only if $f(x)$ does not have a root in k.*

Sketch of Proof. If $f(x) = g(x)h(x)$ and neither g nor h is constant, then $\deg(f) = \deg(g) + \deg(h)$ implies that at least one of the factors has degree 1. •

It is easy to see that Corollary 3.34 can be false if $\deg(f) \geq 4$. For example, consider $f(x) = x^4 + 2x^2 + 1 = (x^2 + 1)^2$ in $\mathbb{R}[x]$.

Example 3.35.

(i) We determine the irreducible polynomials in $\mathbb{I}_2[x]$ of small degree.

As always, the linear polynomials x and $x + 1$ are irreducible.

There are four quadratics: x^2; $x^2 + x$; $x^2 + 1$; $x^2 + x + 1$ (more generally, there are p^n monic polynomials of degree n in $\mathbb{I}_p[x]$, for there are p choices for each of the n coefficients $a_0, \ldots, a_{n-1}$). Since each of the first three has a root in $\mathbb{I}_2$, there is only one irreducible quadratic.

There are eight cubics, of which four are reducible because their constant term is 0. The remaining polynomials are

$$x^3 + 1; \qquad x^3 + x + 1; \qquad x^3 + x^2 + 1; \qquad x^3 + x^2 + x + 1.$$

Since 1 is a root of the first and fourth, the middle two are the only irreducible cubics.

There are 16 quartics, of which eight are reducible because their constant term is 0. Of the eight with nonzero constant term, those having an even number of nonzero coefficients have 1 as a root. There are now only four surviving polynomials $f(x)$, and each of them has no roots in $\mathbb{I}_2$; i.e., they have no linear factors. If $f(x) = g(x)h(x)$, then both $g(x)$ and $h(x)$ must be irreducible quadratics. But there is only one irreducible quadratic, namely, $x^2 + x + 1$, and so $(x^2 + x + 1)^2 = x^4 + x^2 + 1$ is reducible while the other three quartics are irreducible. The following list summarizes these observations.

Irreducible Polynomials of Low Degree over $\mathbb{I}_2$

degree 2:	$x^2 + x + 1$.		
degree 3:	$x^3 + x + 1$;	$x^3 + x^2 + 1$.	
degree 4:	$x^4 + x^3 + 1$;	$x^4 + x + 1$;	$x^4 + x^3 + x^2 + x + 1$.

(ii) Here is a list of the monic irreducible quadratics and cubics in $\mathbb{I}_3[x]$. The reader can verify that the list is correct by first enumerating all such polynomials; there are 6 monic quadratics having nonzero constant term, and there are 18 monic cubics having nonzero constant term. It must then be checked which of these have 1 or -1 as a root (it is more convenient to write -1 instead of 2).

Monic Irreducible Quadratics and Cubics over $\mathbb{I}_3$

degree 2:	$x^2 + 1$;	$x^2 + x - 1$;	$x^2 - x - 1$.
degree 3:	$x^3 - x + 1$;	$x^3 + x^2 - x + 1$;	$x^3 - x^2 + 1$;
	$x^3 - x^2 + x + 1$;	$x^3 - x - 1$;	$x^3 + x^2 - 1$;
	$x^3 + x^2 + x - 1$;	$x^3 - x^2 - x - 1$. ◂	

It is easy to see that if $p(x)$ and $q(x)$ are irreducible polynomials, then $p(x) \mid q(x)$ if and only if there is a unit u with $q(x) = up(x)$. If, in addition, both $p(x)$ and $q(x)$ are monic, then $p(x) \mid q(x)$ implies $p(x) = q(x)$.

Lemma 3.36. *Let k be a field, let $p(x), f(x) \in k[x]$, and let $d(x) = (p, f)$ be their* gcd. *If $p(x)$ is a monic irreducible polynomial, then*

$$d(x) = \begin{cases} 1 & \text{if } p(x) \nmid f(x) \\ p(x) & \text{if } p(x) \mid f(x). \end{cases}$$

Sketch of Proof. Since $d(x) \mid p(x)$, we have $d(x) = 1$ or $d(x) = p(x)$. •

Theorem 3.37 (Euclid's Lemma). *Let k be a field and let $f(x), g(x) \in k[x]$. If $p(x)$ is an irreducible polynomial in $k[x]$, and $p(x) \mid f(x)g(x)$, then either*

$$p(x) \mid f(x) \quad \text{or} \quad p(x) \mid g(x).$$

More generally, if $p(x) \mid f_1(x) \cdots f_n(x)$, then $p(x) \mid f_i(x)$ for some i.

Sketch of Proof. Assume that $p \mid fg$ but that $p \nmid f$. Since p is irreducible, $(p, f) = 1$, and so $1 = sp + tf$ for some polynomials s and t. Therefore,

$$g = spg + tfg.$$

But $p \mid fg$, by hypothesis, and so $p \mid g$. •

Definition. Two polynomials $f(x), g(x) \in k[x]$, where k is a field, are called ***relatively prime*** if their gcd is 1.

Corollary 3.38. *Let $f(x), g(x), h(x) \in k[x]$, where k is a field, and let $h(x)$ and $f(x)$ be relatively prime. If $h(x) \mid f(x)g(x)$, then $h(x) \mid g(x)$.*

Sketch of Proof. The proof of Euclid's lemma also works here: Since $(h, f) = 1$, we have $1 = sh + tf$, and so $g = shg + tfg$. •

Definition. If k is a field, then a rational function $f(x)/g(x) \in k(x)$ is in ***lowest terms*** if $f(x)$ and $g(x)$ are relatively prime.

Proposition 3.39. *If k is a field, every nonzero $f(x)/g(x) \in k(x)$ can be put in lowest terms.*

Sketch of Proof. If $f = df'$ and $g = dg'$, where $d = (f, g)$, then f' and g' are relatively prime, and so f'/g' is in lowest terms. •

The next result allows us to compute gcds.

Theorem 3.40 (Euclidean Algorithm). *If k is a field and $f(x), g(x) \in k[x]$, then there are algorithms for computing the* gcd (f, g), *as well as for finding a pair of polynomials $s(x)$ and $t(x)$ with*

$$(f, g) = s(x)f(x) + t(x)g(x).$$

Proof. The proof is essentially a repetition of the proof of the euclidean algorithm in $\mathbb{Z}$; just iterate the division algorithm:

$$\begin{aligned}
g &= q_1 f + r_1 \\
f &= q_2 r_1 + r_2 \\
r_1 &= q_3 r_2 + r_3 \\
&\;\vdots \\
r_{n-4} &= q_{n-2} r_{n-3} + r_{n-2} \\
r_{n-3} &= q_{n-1} r_{n-2} + r_{n-1} \\
r_{n-2} &= q_n r_{n-1} + r_n \\
r_{n-1} &= q_{n+1} r_n.
\end{aligned}$$

Since the degrees of the remainders are strictly decreasing, this procedure must stop after a finite number of steps. The claim is that $d = r_n$ is the gcd, once it is made monic. We see that d is a common divisor of f and g by back substitution: work from the bottom up. To see that d is the gcd, work from the top down to show that if c is any common divisor of f and g, then $c \mid r_i$ for every i. Finally, to find s and t with $d = sf + tg$, again work

from the bottom up.

$$\begin{aligned} r_n &= r_{n-2} - q_n r_{n-1} \\ &= r_{n-2} - q_n(r_{n-3} - q_{n-1}r_{n-2}) \\ &= (1 + q_{n-1})r_{n-2} - q_n r_{n-3} \\ &= (1 + q_{n-1})(r_{n-4} - q_{n-2}r_{n-3}) - q_n r_{n-3} \\ &= (1 + q_{n-1})r_{n-4} - [(1 + q_{n-1})q_{n-2} + q_n]r_{n-3} \\ &\ \ \vdots \\ &= sf + tg \quad \bullet \end{aligned}$$

Here is an unexpected bonus from the euclidean algorithm.

Corollary 3.41. *Let k be a subfield of a field K, so that $k[x]$ is a subring of $K[x]$. If $f(x), g(x) \in k[x]$, then their* gcd *in $k[x]$ is equal to their* gcd *in $K[x]$.*

Proof. The division algorithm in $K[x]$ gives

$$g(x) = Q(x)f(x) + R(x),$$

where $Q(x), R(x) \in K[x]$; since $f(x), g(x) \in k[x]$, the division algorithm in $k[x]$ gives

$$g(x) = q(x)f(x) + r(x),$$

where $q(x), r(x) \in k[x]$. But the equation $g(x) = q(x)f(x) + r(x)$ also holds in $K[x]$ because $k[x] \subseteq K[x]$, so that the uniqueness of quotient and remainder in the division algorithm in $K[x]$ gives $Q(x) = q(x) \in k[x]$ and $R(x) = r(x) \in k[x]$. Therefore, the list of equations occurring in the euclidean algorithm in $K[x]$ is exactly the same list occurring in the euclidean algorithm in the smaller ring $k[x]$, and so the same gcd is obtained in both polynomial rings. •

For example, the gcd of $x^3 - x^2 + x - 1$ and $x^4 - 1$ is $x^2 + 1$, whether computed in $\mathbb{R}[x]$ or in $\mathbb{C}[x]$, in spite of the fact that there are more divisors with complex coefficients.

Here is the analog for polynomials of the fundamental theorem of arithmetic; it shows that irreducible polynomials are "building blocks" of arbitrary polynomials in the same sense that primes are building blocks of arbitrary integers. To avoid long sentences, let us agree that a "product" may have only one factor. Thus, when we say that a polynomial $f(x)$ is a product of irreducibles, we allow the possibility that the product has only one factor, that is, that $f(x)$ is itself irreducible.

Theorem 3.42 (Unique Factorization). *If k is a field, then every polynomial $f(x) \in k[x]$ of degree ≥ 1 is a product of a nonzero constant and monic irreducibles. Moreover, if $f(x)$ has two such factorizations*

$$f(x) = ap_1(x)\cdots p_m(x) \qquad \textit{and} \qquad f(x) = bq_1(x)\cdots q_n(x),$$

that is, a and b are nonzero constants and the p's and q's are monic irreducibles, then $a = b$, $m = n$, and the q's may be reindexed so that $q_i = p_i$ for all i.

Proof. We prove the existence of a factorization for a polynomial $f(x)$ by (the second form of) induction on $\deg(f) \geq 1$. If $\deg(f) = 1$, then $f(x) = ax + c = a(x + a^{-1}c)$. As every linear polynomial, $x + a^{-1}c$ is irreducible, and so it is a product of irreducibles in our present usage of "product." Assume now that $\deg(f) \geq 1$. If $f(x)$ is irreducible and its leading coefficient is a, write $f(x) = a(a^{-1}f(x))$; we are done, for $a^{-1}f(x)$ is monic. If $f(x)$ is not irreducible, then $f(x) = g(x)h(x)$, where $\deg(g) < \deg(f)$ and $\deg(h) < \deg(f)$. By the inductive hypothesis, there are factorizations $g(x) = bp_1(x) \cdots p_m(x)$ and $h(x) = cq_1(x) \cdots q_n(x)$, where the p's and q's are monic irreducibles. It follows that

$$f(x) = (bc)p_1(x) \cdots p_m(x)q_1(x) \cdots q_n(x),$$

as desired.

We now prove, by induction on $M = \max\{m, n\} \geq 1$, that if there is an equation

$$ap_1(x) \cdots p_m(x) = bq_1(x) \cdots q_n(x)$$

in which a and b are nonzero constants and the p's and q's are monic irreducibles, then $a = b$, $m = n$, and the q's may be reindexed so that $q_i = p_i$ for all i. For the base step $M = 1$, the hypothesis gives a polynomial, call it $g(x)$, with $g(x) = ap_1(x) = bq_1(x)$. Now a is the leading coefficient of $g(x)$, because $p_1(x)$ is monic; similarly, b is the leading coefficient of $g(x)$ because $q_1(x)$ is monic. Therefore, $a = b$, and canceling gives $p_1(x) = q_1(x)$. For the inductive step, the given equation shows that $p_m(x) \mid q_1(x) \cdots q_n(x)$. By Euclid's lemma for polynomials, there is some i with $p_m(x) \mid q_i(x)$. But $q_i(x)$, being monic irreducible, has no monic divisors other than 1 and itself, so that $q_i(x) = p_m(x)$. Reindexing, we may assume that $q_n(x) = p_m(x)$. Canceling this factor, we have $ap_1(x) \cdots p_{m-1}(x) = bq_1(x) \cdots q_{n-1}(x)$. By the inductive hypothesis, $a = b$, $m - 1 = n - 1$ (hence $m = n$), and after possible reindexing, $q_i = p_i$ for all i. •

Let k be a field, and assume that there are $a, r_1, \ldots, r_n \in k$ with

$$f(x) = a\prod_{i=1}^{n}(x - r_i).$$

If $r_1, \ldots, r_s$, where $s \leq n$, are the distinct roots of $f(x)$, then collecting terms gives

$$f(x) = a(x - r_1)^{e_1}(x - r_2)^{e_2} \cdots (x - r_s)^{e_s},$$

where the r_j are distinct and $e_j \geq 1$ for all j. We call e_j the ***multiplicity*** of the root r_j. As linear polynomials are always irreducible, unique factorization shows that multiplicities of roots are well-defined.

Although there are some techniques to help decide whether an integer is prime, the general problem is a very difficult one. It is also very difficult to determine whether a polynomial is irreducible, but we now present some useful techniques that frequently work.

We know that if $f(x) \in k[x]$ and r is a root of $f(x)$ in a field k, then there is a factorization $f(x) = (x-r)g(x)$ in $k[x]$, so that $f(x)$ is not irreducible. In Corollary 3.34, we saw that this decides the matter for quadratic and cubic polynomials in $k[x]$: such polynomials are irreducible in $k[x]$ if and only if they have no roots in k. This is no longer true for polynomials of degree ≥ 4.

Theorem 3.43. *Let $f(x) = a_0 + a_1 x + \cdots + a_n x^n \in \mathbb{Z}[x] \subseteq \mathbb{Q}[x]$. Every rational root r of $f(x)$ has the form b/c, where $b \mid a_0$ and $c \mid a_n$.*

Proof. We may assume that $r = b/c$ is in lowest terms, that is, $(b, c) = 1$. Substituting r into $f(x)$ gives

$$0 = f(b/c) = a_0 + a_1 b/c + \cdots + a_n b^n/c^n,$$

and multiplying through by c^n gives

$$0 = a_0 c^n + a_1 b c^{n-1} + \cdots + a_n b^n.$$

Hence, $a_0 c^n = b(-a_1 c^{n-1} - \cdots - a_n b^{n-1})$, that is, $b \mid a_0 c^n$. Since b and c are relatively prime, it follows that b and c^n are relatively prime, and so Euclid's lemma in $\mathbb{Z}$ gives $b \mid a_0$. Similarly, $a_n b^n = c(-a_{n-1} b^{n-1} - \cdots - a_0 c^{n-1})$, $c \mid a_n b^n$, and $c \mid a_n$. •

Definition. A complex number α is called an ***algebraic integer*** if α is a root of a monic $f(x) \in \mathbb{Z}[x]$.

We note that it is crucial, in the definition of algebraic integer, that $f(x) \in \mathbb{Z}[x]$ be monic. Every algebraic number z, that is, every complex number z that is a root of some polynomial $g(x) \in \mathbb{Q}[x]$, is necessarily a root of some polynomial $h(x) \in \mathbb{Z}[x]$; just clear the denominators of the coefficients of $g(x)$.

Of course, every ordinary integer is an algebraic integer. To contrast ordinary integers with more general algebraic integers, elements of $\mathbb{Z}$ may be called ***rational integers***.

Corollary 3.44. *A rational number z that is an algebraic integer must lie in $\mathbb{Z}$. More precisely, if $f(x) \in \mathbb{Z}[x] \subseteq \mathbb{Q}[x]$ is a monic polynomial, then every rational root of $f(x)$ is an integer that divides the constant term.*

Proof. If $f(x) = a_0 + a_1 x + \cdots + a_n x^n$ is monic, then $a_n = 1$, and Theorem 3.43 applies at once. •

For example, consider $f(x) = x^3 + 4x^2 - 2x - 1 \in \mathbb{Q}[x]$. By Corollary 3.34, this cubic is irreducible if and only if it has no rational root. As $f(x)$ is monic, the candidates for rational roots are ± 1, for these are the only divisors of -1 in $\mathbb{Z}$. But $f(1) = 2$ and $f(-1) = 4$, so that neither 1 nor -1 is a root. Thus, $f(x)$ has no roots in $\mathbb{Q}$, and hence $f(x)$ is irreducible in $\mathbb{Q}[x]$.

This corollary gives a new solution of Exercise 1.15(i) on page 12. If m is an integer that is not a perfect square, then the polynomial $x^2 - m$ has no integer roots, and so $\sqrt{m}$ is irrational. Indeed, the reader can now generalize to nth roots: If m is not an nth power of an integer, then $\sqrt[n]{m}$ is irrational, for any rational root of $x^n - m$ must be an integer.

EXERCISES

3.28 Find the gcd of $x^2 - x - 2$ and $x^3 - 7x + 6$ in $\mathbb{I}_5[x]$, and express it as a linear combination of them.
Hint. The answer is $x - 2$.

3.29 Let R be a domain. If $f(x) \in R[x]$ has degree n, prove that $f(x)$ has at most n roots in R.
Hint. Use $\text{Frac}(R)$.

3.30 Show that the following pseudocode implements the euclidean algorithm finding the gcd $f(x)$ and $g(x)$ in $\mathbb{I}_3[x]$, where $f(x) = x^2 + 1$ and $g(x) = x^3 + x + 1$.

```
Input: g, f
Output: d
d := f; s := g
WHILE s ≠ 0 DO
   rem := remainder(h, s)
   h := s
   s := rem
END WHILE
```

3.31 Prove the converse of Euclid's lemma. Let k be a field and let $f(x) \in k[x]$ be a polynomial of degree ≥ 1; if, whenever $f(x)$ divides a product of two polynomials, it necessarily divides one of the factors, then $f(x)$ is irreducible.

3.32 Let $f(x), g(x) \in R[x]$, where R is a domain. If the leading coefficient of $f(x)$ is a unit in R, then the division algorithm gives a quotient $q(x)$ and a remainder $r(x)$ after dividing $g(x)$ by $f(x)$. Prove that $q(x)$ and $r(x)$ are uniquely determined by $g(x)$ and $f(x)$.
Hint. Use $\text{Frac}(R)$.

3.33 Let k be a field, and let $f(x), g(x) \in k[x]$ be relatively prime. If $h(x) \in k[x]$, prove that $f(x) \mid h(x)$ and $g(x) \mid h(x)$ imply $f(x)g(x) \mid h(x)$.
Hint. See Exercise 1.19 on page 13.

3.34 If k is a field, prove that $\sqrt{1 - x^2} \notin k(x)$, where $k(x)$ is the field of rational functions.
Hint. Mimic a proof that $\sqrt{2}$ is irrational.

3.35 (i) In $R[x]$, where R is a field, let $f = p_1^{e_1} \cdots p_m^{e_m}$ and $g = p_1^{\varepsilon_1} \cdots p_m^{\varepsilon_m}$, where the p_i's are distinct monic irreducibles and $e_i, \varepsilon_i \geq 0$ for all i (as with integers, the device of allowing zero exponents allows us to have the same irreducible factors in the two factorizations). Prove that $f \mid g$ if and only if $e_i \leq \varepsilon_i$ for all i.

(ii) Use the (unique) factorization into irreducibles to give formulas for the gcd and lcm of two polynomials analogous to the formulas in Proposition 1.17.

3.36 If p is a prime, prove that there are exactly $\frac{1}{3}(p^3 - p)$ monic irreducible cubic polynomials in $\mathbb{I}_p[x]$. (A formula for the number of monic irreducible polynomials of degree n in $\mathbb{I}_p[x]$ is given on page 194.)

3.37 (i) Let $f(x) = (x - a_1) \cdots (x - a_n) \in k[x]$, where k is a field. Show that $f(x)$ has ***no repeated roots*** (that is, all the a_i are distinct elements of k) if and only if the gcd $(f, f') = 1$, where $f'(x)$ is the derivative of f.
Hint. Use Exercise 3.24 on page 130.

(ii) Prove that if $p(x) \in \mathbb{Q}[x]$ is an irreducible polynomial, then $p(x)$ has no repeated roots in $\mathbb{C}$.

Hint. Corollary 3.41.

3.38 Let $\zeta = e^{2\pi i/n}$.

(i) Prove that

$$x^n - 1 = (x-1)(x-\zeta)(x-\zeta^2)\cdots(x-\zeta^{n-1})$$

and, if n is odd, that

$$x^n + 1 = (x+1)(x+\zeta)(x+\zeta^2)\cdots(x+\zeta^{n-1}).$$

Hint. Use Corollary 3.29.

(ii) For numbers a and b, prove that

$$a^n - b^n = (a-b)(a-\zeta b)(a-\zeta^2 b)\cdots(a-\zeta^{n-1}b)$$

and, if n is odd, that

$$a^n + b^n = (a+b)(a+\zeta b)(a+\zeta^2 b)\cdots(a+\zeta^{n-1}b).$$

Hint. Set $x = a/b$ if $b \neq 0$.

3.5 HOMOMORPHISMS

Just as homomorphisms are used to compare groups, so are homomorphisms used to compare commutative rings.

Definition. If A and R are (commutative) rings, a (***ring***) ***homomorphism*** is a function $f \colon A \to R$ such that

(i) $f(1) = 1$;

(ii) $f(a + a') = f(a) + f(a')$ for all $a, a' \in A$;

(iii) $f(aa') = f(a)f(a')$ for all $a, a' \in A$.

A homomorphism that is also a bijection is called an ***isomorphism***. Commutative rings A and R are called ***isomorphic***, denoted by $A \cong R$, if there is an isomorphism $f \colon A \to R$.

Example 3.45.

(i) Let R be a domain and let $F = \operatorname{Frac}(R)$ denote its fraction field. In Theorem 3.13 we said that R is a subring of F, but that is not the truth; R is not even a subset of F. We did find a subring R' of F, however, that has a very strong resemblance to R, namely, $R' = \{[a, 1] : a \in R\} \subseteq F$. The function $f \colon R \to R'$, given by $f(a) = [a, 1]$, is easily seen to be an isomorphism.

(ii) When an element in a commutative ring R was "identified" with a constant polynomial [in the proof of Lemma 3.16(iii)], that is, r was identified with $(r, 0, 0, \ldots)$, we implied that R is a subring of $R[x]$. The subset $R' = \{(r, 0, 0, \ldots) : r \in R\}$ is a subring of $R[x]$,

and it is easy to see that the function $f: R \to R'$, defined by $f(r) = (r, 0, 0, \dots)$, is an isomorphism.

(iii) If S is a subring of a commutative ring R, then the inclusion $i: S \to R$ is a ring homomorphism because we have insisted that the identity 1 of R lies in S. [See Exercise 3.4(iii) on page 124.] ◀

Example 3.46.
(i) Complex conjugation $z = a + ib \mapsto \overline{z} = a - ib$ is an isomorphism $\mathbb{C} \to \mathbb{C}$ because $\overline{1} = 1$, $\overline{z + w} = \overline{z} + \overline{w}$, and $\overline{zw} = \overline{z}\,\overline{w}$

(ii) Here is an example of a homomorphism of rings that is not an isomorphism. Choose $m \geq 2$ and define $f: \mathbb{Z} \to \mathbb{I}_m$ by $f(n) = [n]$. Notice that f is surjective (but not injective).

(iii) The preceding example can be generalized. If R is a commutative ring with its "one" denoted by ε, then the function $\chi: \mathbb{Z} \to R$, defined by $\chi(n) = n\varepsilon$, is a ring homomorphism.[8]

(iv) Let R be a commutative ring, and let $a \in R$. Define the ***evaluation homomorphism*** $e_a: R[x] \to R$ by $e_a(f(x)) = f(a)$; that is, if $f(x) = \sum r_i x^i$, then $f(a) = \sum r_i a^i$. We let the reader check that e_a is a ring homomorphism. ◀

Certain properties of a ring homomorphism $f: A \to R$ follow from its being a homomorphism between the additive groups A and R. For example, $f(0) = 0$, $f(-a) = -f(a)$, and $f(na) = nf(a)$ for all $n \in \mathbb{Z}$.

Lemma 3.47. *If $f: A \to R$ is a ring homomorphism, then, for all $a \in A$,*

(i) *$f(a^n) = f(a)^n$ for all $n \geq 0$;*

(ii) *if a is a unit, then $f(a)$ is a unit and $f(a^{-1}) = f(a)^{-1}$; in fact, if a is a unit, then $f(a^{-n}) = f(a)^{-n}$ for all $n \geq 1$;*

(iii) *if $f: A \to R$ is a ring homomorphism, then*

$$f(U(A)) \leq U(R),$$

where $U(A)$ is the group of units of A; if f is an isomorphism, then

$$U(A) \cong U(R).$$

Sketch of Proof. (i) Induction on $n \geq 0$.

(ii) If $ab = 1$, then $1 = f(ab) = f(a)f(b)$. The last statement follows by induction on $n \geq 1$.

(iii) Immediate, from part (ii). •

[8] Recall that if $a \in R$ and n is a positive integer, then na is the additive version of the multiplicative notation a^n; that is, na is the sum of a with itself n times.

Proposition 3.48. *If R and S are commutative rings and $\varphi\colon R \to S$ is a ring homomorphism, then there is a ring homomorphism $\varphi^*\colon R[x] \to S[x]$ given by*

$$\varphi^*\colon r_0 + r_1 x + r_2 x^2 + \cdots \mapsto \varphi(r_0) + \varphi(r_1)x + \varphi(r_2)x^2 + \cdots .$$

Sketch of Proof. It is clear that φ^* is well-defined, and a routine calculation shows that it is a ring homomorphism. •

Definition. If $f\colon A \to R$ is a ring homomorphism, then its ***kernel*** is

$$\textbf{ker}\, \boldsymbol{f} = \{a \in A \text{ with } f(a) = 0\},$$

and its ***image*** is

$$\textbf{im}\, \boldsymbol{f} = \{r \in R : r = f(a) \text{ for some } a \in R\}.$$

Notice that if we forget their multiplications, then the rings A and R are additive abelian groups and these definitions coincide with the group-theoretic ones.

Let k be a commutative ring, let $a \in k$, and, as in Example 3.46(iv), consider the evaluation homomorphism $e_a : k[x] \to k$ sending $f(x) \mapsto f(a)$. Now e_a is always surjective, for if $b \in k$, then $b = e_a(f)$, where $f(x) = x - a + b$. By definition, $\ker e_a$ consists of all those polynomials $g(x)$ for which $g(a) = 0$; that is, $\ker e_a$ consists of all the polynomials in $k[x]$ having a as a root.

The kernel of a group homomorphism is not merely a subgroup; it is a *normal* subgroup; that is, it is also closed under conjugation by any element in the ambient group. Similarly, if R is not the zero ring, the kernel of a ring homomorphism $f\colon A \to R$ is almost a subring [$\ker f$ is not a subring because it never contains 1: $f(1) = 1 \neq 0$], and we shall see that it is closed under multiplication by any element in the ambient ring.

Definition. An ***ideal*** in a commutative ring R is a subset I of R such that

(i) $0 \in I$;

(ii) if $a, b \in I$, then $a + b \in I$;[9]

(iii) if $a \in I$ and $r \in R$, then $ra \in I$.

The ring R itself and the subset consisting of 0 alone, which we denote by $\{0\}$, are always ideals in a commutative ring R. An ideal $I \neq R$ is called a ***proper ideal***.

Example 3.49.
If $b_1, b_2, \ldots, b_n$ lie in R, then the set of all linear combinations

$$I = \left\{r_1 b_1 + r_2 b_2 + \cdots + r_n b_n : r_i \in R \text{ for all } i\right\}$$

[9]In contrast to the definition of subring, it suffices to assume that $a + b \in I$ instead of $a - b \in I$. If I is an ideal and $b \in I$, then $(-1)b \in I$, and so $a - b = a + (-1)b \in I$.

is an ideal in R. We write $I = (b_1, b_2, \ldots, b_n)$ in this case, and we call I the ***ideal generated by*** $b_1, b_2, \ldots, b_n$. In particular, if $n = 1$, then

$$I = (b) = \{rb : r \in R\}$$

is an ideal in R; (b) consists of all the multiples of b, and it is called the ***principal ideal*** generated by b. Notice that R and $\{0\}$ are always principal ideals: $R = (1)$ and $\{0\} = (0)$. In $\mathbb{Z}$, the even integers form the principal ideal (2). ◀

Proposition 3.50. *If $f : A \to R$ is a ring homomorphism, then* $\ker f$ *is an ideal in A and* $\operatorname{im} f$ *is a subring of R. Moreover, if A and R are not zero rings, then* $\ker f$ *is a proper ideal.*

Sketch of Proof. $\ker f$ is an additive subgroup of A; moreover, if $u \in \ker f$ and $a \in A$, then $f(au) = f(a)f(u) = f(a) \cdot 0 = 0$. Hence, $\ker f$ is an ideal. If R is not the zero ring, then $1 \neq 0$; hence, the identity $1 \in A$ does not lie in $\ker f$, because $f(1) = 1 \neq 0$ in R, and so $\ker f$ is a proper ideal. It is routine to check that $\operatorname{im} f$ is a subring of R. •

Example 3.51.
(i) If an ideal I in a commutative ring R contains 1, then $I = R$, for now I contains $r = r1$ for every $r \in R$. Indeed, if I contains a unit u, then $I = R$, for then I contains $u^{-1}u = 1$.

(ii) It follows from (i) that if R is a field, then the only ideals I in R are $\{0\}$ and R itself: if $I \neq \{0\}$, it contains some nonzero element, and every nonzero element in a field is a unit.

Conversely, assume that R is a nonzero commutative ring whose only ideals are R itself and $\{0\}$. If $a \in R$ and $a \neq 0$, then the principal ideal $(a) = R$, for $(a) \neq 0$, and so $1 \in R = (a)$. There is thus $r \in R$ with $1 = ra$; that is, a has an inverse in R, and so R is a field. ◀

Proposition 3.52. *A ring homomorphism $f : A \to R$ is an injection if and only if* $\ker f = \{0\}$.

Sketch of Proof. This follows from the corresponding result for group homomorphisms, because f is a homomorphism from the additive group of A to the additive group of R. •

Corollary 3.53. *If $f : k \to R$ is a ring homomorphism, where k is a field and R is not the zero ring, then f is an injection.*

Proof. The only proper ideal in k is $\{0\}$. •

Theorem 3.54. *If k is a field, then every ideal I in $k[x]$ is a principal ideal. Moreover, if $I \neq \{0\}$, there is a monic polynomial that generates I.*

Sketch of Proof. If k is a field, then $k[x]$ is an example of a *euclidean ring*. In Theorem 3.60, we will prove that every ideal in a euclidean ring is a principal ideal. •

Definition. A domain R is a ***principal ideal domain*** if every ideal in R is a principal ideal. This name is often abbreviated to PID.

Example 3.55.
(i) The ring of integers is a PID.

(ii) Every field is a PID, by Example 3.51(ii).

(iii) If k is a field, then the polynomial ring $k[x]$ is a PID, by Theorem 3.54.

(iv) There are rings other than $\mathbb{Z}$ and $k[x]$, where k is a field, that have a division algorithm; they are called *euclidean rings*, and they, too, are PIDs. We shall consider them in the next section. ◄

It is not true that ideals in arbitrary commutative rings are always principal ideals.

Example 3.56.
Let $R = \mathbb{Z}[x]$, the commutative ring of all polynomials over $\mathbb{Z}$. It is easy to see that the set I of all polynomials with even constant term is an ideal in $\mathbb{Z}[x]$. We show that I is *not* a principal ideal.

Suppose there is $d(x) \in \mathbb{Z}[x]$ with $I = (d(x))$. The constant $2 \in I$, so that there is $f(x) \in \mathbb{Z}[x]$ with $2 = d(x)f(x)$. Since the degree of a product is the sum of the degrees of the factors, $0 = \deg(2) = \deg(d) + \deg(f)$. Since degrees are nonnegative, it follows that $\deg(d) = 0$ [i.e., $d(x)$ is a nonzero constant]. As constants here are integers, the candidates for $d(x)$ are ± 1 and ± 2. Suppose $d(x) = \pm 2$; since $x \in I$, there is $g(x) \in \mathbb{Z}[x]$ with $x = d(x)g(x) = \pm 2g(x)$. But every coefficient on the right side is even, while the coefficient of x on the left side is 1. This contradiction gives $d(x) = \pm 1$. By Example 3.51(ii), $I = \mathbb{Z}[x]$, another contradiction. Therefore, no such $d(x)$ exists, that is, the ideal I is not a principal ideal. ◄

Certain theorems holding in $\mathbb{Z}$ carry over to PIDs once the standard definitions are generalized; the notion of divisor has already been generalized.

Definition. An element δ in a commutative ring R is a ***greatest common divisor***, gcd, of elements $\alpha, \beta \in R$ if

(i) δ is a common divisor of α and β;

(ii) if γ is any common divisor of α and β, then $\gamma \mid \delta$.

Greatest common divisors, when they exist, need not be unique; for example, it is easy to see that if c is a greatest common divisor of f and g, then so is uc for any unit $u \in R$. In the special case $R = \mathbb{Z}$, we force uniqueness of the gcd by requiring it to be positive; if $R = k[x]$, where k is a field, then we force uniqueness of the gcd by further requiring it to be monic.

Remark. Let R be a PID and let $\pi, \alpha \in R$ with π irreducible. A gcd δ of π and α is, in particular, a divisor of π. Hence, $\pi = \delta\varepsilon$, and irreducibility of π forces either δ or ε to be a unit. Now $\alpha = \delta\beta$. If δ is not a unit, then ε is a unit, and so

$$\alpha = \delta\beta = \pi\varepsilon^{-1}\beta;$$

that is, $\pi \mid \alpha$. We conclude that if $\pi \nmid \alpha$, then δ is a unit; that is, 1 is a gcd of π and α. ◄

For an example of a domain in which a pair of elements does not have a gcd, see Exercise 3.60 on page 158.

Theorem 3.57. *Let R be a* PID.

(i) *Every $\alpha, \beta \in R$ has a* gcd, *δ, which is a linear combination of α and β:*

$$\delta = \sigma\alpha + \tau\beta,$$

where $\sigma, \tau \in R$.

(ii) *If an irreducible element $\pi \in R$ divides a product $\alpha\beta$, then either $\pi \mid \alpha$ or $\pi \mid \beta$.*

Proof. (i) We may assume that at least one of α and β is not zero (otherwise, the gcd is 0 and the result is obvious). Consider the set I of all the linear combinations:

$$I = \{\sigma\alpha + \tau\beta : \sigma, \tau \text{ in } R\}.$$

Now α and β are in I (take $\sigma = 1$ and $\tau = 0$ or vice versa). It is easy to check that I is an ideal in R, and so there is $\delta \in I$ with $I = (\delta)$, because R is a PID; we claim that δ is a gcd of α and β.

Since $\alpha \in I = (\delta)$, we have $\alpha = \rho\delta$ for some $\rho \in R$; that is, δ is a divisor of α; similarly, δ is a divisor of β, and so δ is a common divisor of α and β.

Since $\delta \in I$, it is a linear combination of α and β: There are $\sigma, \tau \in R$ with

$$\delta = \sigma\alpha + \tau\beta.$$

Finally, if γ is any common divisor of α and β, then $\alpha = \gamma\alpha'$ and $\beta = \gamma\beta'$, so that γ divides δ, for $\delta = \sigma\alpha + \tau\beta = \gamma(\sigma\alpha' + \tau\beta')$. We conclude that δ is a gcd.

(ii) If $\pi \mid \alpha$, we are done. If $\pi \nmid \alpha$, then the remark says that 1 is a gcd of π and α. There are thus elements $\sigma, \tau \in R$ with $1 = \sigma\pi + \tau\alpha$, and so

$$\beta = \sigma\pi\beta + \tau\alpha\beta.$$

Since $\pi \mid \alpha\beta$, it follows that $\pi \mid \beta$, as desired. •

Example 3.58.
If I and J are ideals in a commutative ring R, we now show that $I \cap J$ is also an ideal in R. Since $0 \in I$ and $0 \in J$, we have $0 \in I \cap J$. If $a, b \in I \cap J$, then $a - b \in I$ and $a - b \in J$, for each is an ideal, and so $a - b \in I \cap J$. If $a \in I \cap J$ and $r \in R$, then $ra \in I$ and $ra \in J$, hence $ra \in I \cap J$. Therefore, $I \cap J$ is an ideal. With minor alterations, this argument also proves that the intersection of any family of ideals in R is also an ideal in R. ◄

Definition. If f and g are elements in a commutative ring R, then a ***common multiple*** is an element $m \in R$ with $f \mid m$ and $g \mid m$. If f and g in R are not both 0, define their ***least common multiple***, abbreviated lcm, to be a common multiple c of them with $c \mid m$ for every common multiple m. If $f = 0 = g$, define their lcm $= 0$. The lcm of f and g is often denoted by $[f, g]$.

Least common multiples, when they exist, need not be unique; for example, it is easy to see that if c is a least common multiple of f and g, then so is uc for any unit $u \in R$. In the special case $R = \mathbb{Z}$, we force uniqueness of the lcm by requiring it to be positive; if $R = k[x]$, where k is a field, then we force uniqueness of the lcm by further requiring it to be monic.

EXERCISES

3.39 (i) Let $\varphi\colon A \to R$ be an isomorphism, and let $\psi\colon R \to A$ be its inverse. Show that ψ is an isomorphism.

(ii) Show that the composite of two homomorphisms (isomorphisms) is again a homomorphism (isomorphism).

(iii) Show that $A \cong R$ defines an equivalence relation on the class of all commutative rings.

3.40 Let R be a commutative ring and let $\mathcal{F}(R)$ be the commutative ring of all functions $f\colon R \to R$ with pointwise operations.

(i) Show that R is isomorphic to the subring of $\mathcal{F}(R)$ consisting of all the constant functions.

(ii) If $f(x) \in R[x]$, let $\varphi_f\colon R \to R$ be defined by $r \mapsto f(r)$ [thus, φ_f is the polynomial function associated to $f(x)$]. Show that the function $\varphi\colon R[x] \to \mathcal{F}(R)$, defined by $\varphi(f(x)) = \varphi_f$, is a ring homomorphism.

(iii) Show that φ is injective if R is an infinite field.

3.41 Let I and J be nonzero ideals in a commutative ring R. If R is a domain, prove that $I \cap J \neq \{0\}$.

3.42 Let R be a commutative ring. Show that the function $\varepsilon\colon R[x] \to R$, defined by

$$\varepsilon\colon a_0 + a_1x + a_2x + \cdots + a_nx^n \mapsto a_0,$$

is a homomorphism. Describe $\ker\varepsilon$ in terms of roots of polynomials.

3.43 If R is a commutative ring and $c \in R$, prove that the function $\varphi\colon R[x] \to R[x]$, defined by $f(x) \mapsto f(x+c)$, is an isomorphism. In more detail, $\varphi(\sum_i s_ix^i) = \sum_i s_i(x+c)^i$.
Hint. This is a routine but long calculation.

3.44 (i) Prove that F, the field with four elements (see Exercise 3.14 on page 125), and $\mathbb{I}_4$ are not isomorphic commutative rings.

(ii) Prove that any two fields having exactly four elements are isomorphic.

Hint. First prove that $1+1 = 0$, and then show that the nonzero elements form a cyclic group of order 3 under multiplication.

3.45 (i) Show that every element $a \in \mathbb{I}_p$ has a pth root (i.e., there is $b \in \mathbb{I}_p$ with $a = b^p$).

(ii) Let k be a field that contains $\mathbb{I}_p$ as a subfield [e.g., $k = \mathbb{I}_p(x)$]. For every positive integer n, show that the function $\varphi_n\colon k \to k$, given by $\varphi(a) = a^{p^n}$, is a ring homomorphism.

3.46 If R is a field, show that $R \cong \mathrm{Frac}(R)$. More precisely, show that the homomorphism $f : R \to \mathrm{Frac}(R)$ in Example 3.45(i), namely, $r \mapsto [r, 1]$, is an isomorphism.

3.47 (i) If A and R are domains and $\varphi : A \to R$ is a ring isomorphism, prove that

$$[a, b] \mapsto [\varphi(a), \varphi(b)]$$

is a ring isomorphism $\mathrm{Frac}(A) \to \mathrm{Frac}(R)$.

(ii) Prove that if a field k contains an isomorphic copy of $\mathbb{Z}$ as a subring, then k must contain an isomorphic copy of $\mathbb{Q}$.

(iii) Let R be a domain and let $\varphi : R \to k$ be an injective ring homomorphism, where k is a field. Prove that there exists a unique ring homomorphism $\Phi : \mathrm{Frac}(R) \to k$ extending φ; that is, $\Phi|R = \varphi$.

3.48 Let R be a domain with fraction field $F = \mathrm{Frac}(R)$.

(i) Prove that $\mathrm{Frac}(R[x]) \cong F(x)$.

(ii) Prove that $\mathrm{Frac}(R[x_1, x_2, \ldots, x_n]) \cong F(x_1, x_2, \ldots, x_n)$ (see page 129).

3.49 (i) If R and S are commutative rings, show that their ***direct product*** $R \times S$ is also a commutative ring, where addition and multiplication in $R \times S$ are defined "coordinatewise":

$$(r, s) + (r', s') = (r + r', s + s') \quad \text{and} \quad (r, s)(r', s') = (rr', ss').$$

(ii) Show that if m and n are relatively prime, then $\mathbb{I}_{mn} \cong \mathbb{I}_m \times \mathbb{I}_n$ as rings.

Hint. See Theorem 2.81.

(iii) Show that if neither R nor S is the zero ring, then $R \times S$ is not a domain.

(iv) Show that $R \times \{0\}$ is an ideal in $R \times S$.

(v) Show that $R \times \{0\}$ is a ring isomorphic to R, but it is not a subring of $R \times S$.

3.50 (i) If R and S are nonzero commutative rings, prove that

$$U(R \times S) = U(R) \times U(S),$$

where $U(R)$ is the group of units of R.

Hint. Show that (r, s) is a unit in $R \times S$ if and only if r is a unit in R and s is a unit in S.

(ii) Redo Exercise 2.65 on page 94 using part (i).

(iii) Use part (i) to give another proof of Corollary 2.83.

3.51 Let F be the set of all 2×2 real matrices of the form

$$A = \begin{bmatrix} a & b \\ -b & a \end{bmatrix}.$$

Prove that F is a field (with operations matrix addition and matrix multiplication), and prove that there is an isomorphism $\varphi : F \to \mathbb{C}$ with $\det(A) = \varphi(A)\overline{\varphi(A)}$.

Hint. Define $\varphi : F \to \mathbb{C}$ by $\varphi(A) = a + ib$.

3.52 If k is a field and $[f, g]$ denotes the lcm of monic polynomials $f(x), g(x) \in k[x]$, show that

$$[f, g](f, g) = fg.$$

Hint. See Exercise 1.26 on page 13. By definition, lcm's are monic.

3.53 If R is a PID and $a, b \in R$, prove that their lcm exists.

3.54 (i) If k is a field, prove that the ring of formal power series $k[[x]]$ is a PID.

Hint. If I is a nonzero ideal, choose $\tau \in I$ of smallest order. Use Exercise 3.27 on page 130 to prove that $I = (\tau)$.

(ii) Prove that every nonzero ideal in $k[[x]]$ is equal to (x^n) for some $n \geq 0$.

3.55 If k is a field, show that the ideal (x, y) in $k[x, y]$ is not a principal ideal (see page 129).

3.56 For every $m \geq 1$, prove that every ideal in $\mathbb{I}_m$ is a principal ideal. (If m is composite, then $\mathbb{I}_m$ is not a PID because it is not a domain.)

3.6 EUCLIDEAN RINGS

There are rings other than $\mathbb{Z}$ and $k[x]$, where k is a field, that have a division algorithm. In particular, we present an example of such a ring in which the quotient and remainder are not unique. We begin by generalizing a property shared by both $\mathbb{Z}$ and $k[x]$.

Definition. A ***euclidean ring*** is a domain R that is equipped with a function

$$\partial : R - \{0\} \to \mathbb{N},$$

called a ***degree function***, such that

(i) $\partial(f) \leq \partial(fg)$ for all $f, g \in R$ with $f, g \neq 0$;

(ii) for all $f, g \in R$ with $f \neq 0$, there exist $q, r \in R$ with

$$g = qf + r,$$

where either $r = 0$ or $\partial(r) < \partial(f)$.

Note that if R has a degree function ∂ that is identically 0, then condition (ii) forces $r = 0$ always; taking $g = 1$ shows that R is a field in this case.

Example 3.59.
(i) The integers $\mathbb{Z}$ is a euclidean ring with degree function $\partial(m) = |m|$. In $\mathbb{Z}$, we have

$$\partial(mn) = |mn| = |m||n| = \partial(m)\partial(n).$$

(ii) When k is a field, the domain $k[x]$ is a euclidean ring with degree function the usual degree of a nonzero polynomial. In $k[x]$, we have

$$\begin{aligned}\partial(fg) &= \deg(fg)\\ &= \deg(f) + \deg(g)\\ &= \partial(f) + \partial(g).\end{aligned}$$

Since $\partial(mn) = \partial(m)\partial(n)$ in $\mathbb{Z}$, the behavior of the degree of a product is not determined by the axioms in the definition of a degree function. If a degree function ∂ is multiplicative, that is, if

$$\partial(fg) = \partial(f)\partial(g),$$

then ∂ is called a ***norm***.

(iii) The Gaussian[10] integers $\mathbb{Z}[i]$ form a euclidean ring whose degree function

$$\partial(a + bi) = a^2 + b^2$$

is a norm. One reason for showing that $\mathbb{Z}[i]$ is a euclidean ring is that it is then a PID, and hence it has unique factorization of its elements into products of irreducibles; Gauss used this fact in his proof that if an odd prime p is sum of two squares, say $p = a^2 + b^2$, where a and b are natural numbers, then the pair a, b is unique (see Theorem 3.66).

To see that ∂ is a multiplicative degree function, note first that if $\alpha = a + bi$, then

$$\partial(\alpha) = \alpha\overline{\alpha},$$

where $\overline{\alpha} = a - bi$ is the complex conjugate of α. It follows that $\partial(\alpha\beta) = \partial(\alpha)\partial(\beta)$ for all $\alpha, \beta \in \mathbb{Z}[i]$, because

$$\partial(\alpha\beta) = \alpha\beta\overline{\alpha\beta} = \alpha\beta\overline{\alpha}\overline{\beta} = \alpha\overline{\alpha}\beta\overline{\beta} = \partial(\alpha)\partial(\beta);$$

indeed, this is even true for all $\alpha, \beta \in \mathbb{Q}[i] = \{x + yi : x, y \in \mathbb{Q}\}$, by Corollary 1.31.

We now show that ∂ satisfies the first property of a degree function. If $\beta = c + id \in \mathbb{Z}[i]$ and $\beta \neq 0$, then

$$1 \leq \partial(\beta),$$

for $\partial(\beta) = c^2 + d^2$ is a positive integer; it follows that if $\alpha, \beta \in \mathbb{Z}[i]$ and $\beta \neq 0$, then

$$\partial(\alpha) \leq \partial(\alpha)\partial(\beta) = \partial(\alpha\beta).$$

Let us show that ∂ also satisfies the second desired property. Given $\alpha, \beta \in \mathbb{Z}[i]$ with $\beta \neq 0$, regard α/β as an element of $\mathbb{C}$. Rationalizing the denominator gives $\alpha/\beta = \alpha\overline{\beta}/\beta\overline{\beta} = \alpha\overline{\beta}/\partial(\beta)$, so that

$$\alpha/\beta = x + yi,$$

where $x, y \in \mathbb{Q}$. Write $x = a + u$ and $y = b + v$, where $a, b \in \mathbb{Z}$ are integers closest to x and y, respectively; thus, $|u|, |v| \leq \frac{1}{2}$. (If x or y has the form $m + \frac{1}{2}$, where m is an integer, then there is a choice of nearest integer: $x = m + \frac{1}{2}$ or $x = (m + 1) - \frac{1}{2}$; a similar choice arises if x or y has the form $m - \frac{1}{2}$.) It follows that

$$\alpha = \beta(a + bi) + \beta(u + vi).$$

[10]The Gaussian integers are so called because Gauss tacitly used $\mathbb{Z}[i]$ and its norm ∂ to investigate biquadratic residues.

Notice that $\beta(u+vi) \in \mathbb{Z}[i]$, for it is equal to $\alpha - \beta(a+bi)$. Finally, we have

$$\partial\big(\beta(u+vi)\big) = \partial(\beta)\partial(u+vi),$$

and so ∂ will be a degree function if $\partial(u+vi) < 1$. And this is so, for the inequalities $|u| \le \frac{1}{2}$ and $|v| \le \frac{1}{2}$ give $u^2 \le \frac{1}{4}$ and $v^2 \le \frac{1}{4}$, and hence $\partial(u+vi) = u^2 + v^2 \le \frac{1}{4} + \frac{1}{4} = \frac{1}{2} < 1$. Therefore, $\partial(\beta(u+vi)) < \partial(\beta)$, and so $\mathbb{Z}[i]$ is a euclidean ring whose degree function is a norm.

We now show that quotients and remainders may not be unique (because of the choices noted previously). For example, let $\alpha = 3+5i$ and $\beta = 2$. Then $\alpha/\beta = \frac{3}{2} + \frac{5}{2}i$; the choices are

$$a = 1 \text{ and } u = \tfrac{1}{2} \quad \text{or} \quad a = 2 \text{ and } u = -\tfrac{1}{2};$$
$$b = 2 \text{ and } v = \tfrac{1}{2} \quad \text{or} \quad b = 3 \text{ and } v = -\tfrac{1}{2}.$$

There are four quotients and remainders after dividing $3+5i$ by 2 in $\mathbb{Z}[i]$, for each of the remainders (e.g., $1+i$) has degree $2 < 4 = \partial(2)$:

$$\begin{aligned} 3+5i &= 2(1+2i) + (1+i); \\ &= 2(1+3i) + (1-i); \\ &= 2(2+2i) + (-1+i); \\ &= 2(2+3i) + (-1-i). \quad \blacktriangleleft \end{aligned}$$

Theorem 3.60. *Every euclidean ring R is a* PID.

Proof. Let I be an ideal in R. If $I = \{0\}$, then $I = (0)$ is principal; therefore, we may assume that $I \neq (0)$. By the least integer axiom, the set of all degrees of nonzero elements in I has a smallest element, say, n; choose $d \in I$ with $\partial(d) = n$. Clearly, $(d) \subseteq I$, and so it suffices to prove the reverse inclusion. If $a \in I$, then there are $q, r \in R$ with $a = qd + r$, where either $r = 0$ or $\partial(r) < \partial(d)$. But $r = a - qd \in I$, and so d having least degree implies that $r = 0$. Hence, $a = qd \in (d)$, and $I = (d)$. •

Corollary 3.61. *The ring of Gaussian integers $\mathbb{Z}[i]$ is a principal ideal domain.*

The converse of Theorem 3.60 is false: There are PIDs that are not euclidean rings, as we see in the next example.

Example 3.62.
It is shown in algebraic number theory that the ring

$$R = \{a + b\alpha : a, b \in \mathbb{Z}\},$$

where $\alpha = \frac{1}{2}(1 + \sqrt{-19})$, is a PID [$R$ is the ring of algebraic integers in the quadratic number field $\mathbb{Q}(\sqrt{-19})$]. In 1949, T. S. Motzkin showed that R is not a euclidean ring by showing that it does not have a certain property of euclidean rings that does not mention its degree function.

Definition. An element u in a domain R is a ***universal side divisor*** if u is not a unit and, for every $x \in R$, either $u \mid x$ or there is a unit $z \in R$ with $u \mid (x + z)$.

Proposition 3.63. *If R is a euclidean ring but not a field, then R has a universal side divisor.*

Proof. Define

$$S = \{\partial(v) : v \neq 0 \text{ and } v \text{ is not a unit}\},$$

where ∂ is the degree function on R. Since R is not a field, by hypothesis, S is a nonempty subset of the natural numbers. By the least integer axiom, S has a smallest element, say, $\partial(u)$. We claim that u is a universal side divisor. If $x \in R$, there are elements q and r with $x = qu + r$, where either $r = 0$ or $\partial(r) < \partial(u)$. If $r = 0$, then $u \mid x$; if $r \neq 0$, then r must be a unit, otherwise its existence contradicts $\partial(u)$ being the smallest number in S. We have shown that u is a universal side divisor. •

Motzkin then showed that the ring $\{a + b\alpha : a, b \in \mathbb{Z}\}$, where $\alpha = \frac{1}{2}(1 + \sqrt{-19})$, has no universal side divisors, concluding that this PID is not a euclidean ring. For details, we refer the reader to K. S. Williams, "Note on Non-euclidean Principal Ideal Domains," *Math. Mag.* 48 (1975), 176–177. ◂

What are the units in the Gaussian integers?

Proposition 3.64.

(i) *Let R be a euclidean ring R that is not a field. If the degree function ∂ is a norm, then α is a unit if and only if $\partial(\alpha) = 1$.*

(ii) *Let R be a euclidean ring R that is not a field. If the degree function ∂ is a norm and if $\partial(\alpha) = p$, where p is a prime number, then α is irreducible.*

(iii) *The only units in the ring $\mathbb{Z}[i]$ of Gaussian integers are ± 1 and $\pm i$.*

Proof. (i) Since $1^2 = 1$, we have $\partial(1)^2 = \partial(1)$, so that $\partial(1) = 0$ or $\partial(1) = 1$. If $\partial(1) = 0$, then $\partial(a) = \partial(1a) = \partial(1)\partial(a) = 0$ for all $a \in R$. But R is not a field, and so ∂ is not identically zero. We conclude that $\partial(1) = 1$.

If $\alpha \in R$ is a unit, then there is $\beta \in R$ with $\alpha\beta = 1$. Therefore, $\partial(\alpha)\partial(\beta) = 1$. Since the values of ∂ are nonnegative integers, $\partial(\alpha) = 1$.

For the converse, we begin by showing that there is no element $\beta \in R$ with $\partial(\beta) = 0$. If such an element existed, the division algorithm would give $1 = q\beta + r$, where $q, r \in R$ and either $r = 0$ or $\partial(r) < \partial(\beta) = 0$. The inequality cannot occur, and so $r = 0$; that is, β is a unit. But if β is a unit, then $\partial(\beta) = 1$, as we have just proved, and this contradicts $\partial(\beta) = 0$.

Assume now that $\partial(\alpha) = 1$. The division algorithm gives $q, r \in R$ with

$$\alpha = q\alpha^2 + r,$$

where $r = 0$ or $\partial(r) < \partial(\alpha^2)$. As $\partial(\alpha^2) = \partial(\alpha)^2 = 1$, either $r = 0$ or $\partial(r) = 0$. But we have just seen that $\partial(r) = 0$ cannot occur, so that $r = 0$ and $\alpha = q\alpha^2$. It follows that $1 = q\alpha$, and so α is a unit.

(ii) If, on the contrary, $\alpha = \beta\gamma$, where neither β nor γ is a unit, then $p = \partial(\alpha) = \partial(\beta)\partial(\gamma)$. As p is a prime, either $\partial(\beta) = 1$ or $\partial(\gamma) = 1$. By part (i), either β or γ is a unit; that is, α is irreducible.

(iii) If $\alpha = a + bi \in \mathbb{Z}[i]$ is a unit, then $1 = \partial(\alpha) = a^2 + b^2$. This can happen if and only if $a^2 = 1$ and $b^2 = 0$ or $a^2 = 0$ and $b^2 = 1$; that is, $\alpha = \pm 1$ or $\alpha = \pm i$. •

If n is an odd number, then either $n \equiv 1 \bmod 4$ or $n \equiv 3 \bmod 4$; consequently, the odd prime numbers are divided into two classes. For example, 5, 13, 17 are congruent to 1 mod 4, while 3, 7, 11 are congruent to 3 mod 4.

Lemma 3.65. *If p is a prime and $p \equiv 1 \bmod 4$, then there is an integer m with*

$$m^2 \equiv -1 \bmod p.$$

Proof. If $G = (\mathbb{I}_p)^\times$ is the multiplicative group of nonzero elements in $\mathbb{I}_p$, then $|G| = p - 1 \equiv 0 \bmod 4$; that is, 4 is a divisor of $|G|$. By Proposition 2.78, G contains a subgroup S of order 4. By Exercise 2.36 on page 72, either S is cyclic or $a^2 = 1$ for all $a \in S$. Since $\mathbb{I}_p$ is a field, however, it cannot contain four roots of the quadratic $x^2 - 1$. Therefore, S is cyclic,[11] say, $S = \langle [m] \rangle$, where $[m]$ is the congruence class of m mod p. Since $[m]$ has order 4, we have $[m^4] = [1]$. Moreover, $[m^2] \neq [1]$ (lest $[m]$ have order $\leq 2 < 4$), and so $[m^2] = [-1]$, for $[-1]$ is the unique element in S of order 2. Therefore, $m^2 \equiv -1 \bmod p$. •

Theorem 3.66 (Fermat's[12] Two-Squares Theorem). *An odd prime p is a sum of two squares,*

$$p = a^2 + b^2,$$

where a and b are integers, if and only if $p \equiv 1 \bmod 4$.

Proof. Assume that $p = a^2 + b^2$. Since p is odd, a and b have different parity; say, a is even and b is odd. Hence, $a = 2m$ and $b = 2n + 1$, and

$$p = a^2 + b^2 = 4m^2 + 4n^2 + 4n + 1 \equiv 1 \bmod 4.$$

Conversely, assume that $p \equiv 1 \bmod 4$. By the lemma, there is an integer m such that

$$p \mid (m^2 + 1).$$

[11] Theorem 3.30 says that G is a cyclic group, which implies that S is cyclic, for every subgroup of a cyclic group is itself cyclic. We choose to avoid this theorem here, for the proof just given is more elementary.

[12] Fermat was the first to state this theorem, but the first published proof is due to Euler. Gauss proved that there is only one pair of natural numbers a and b with $p = a^2 + b^2$.

In $\mathbb{Z}[i]$, there is a factorization $m^2 + 1 = (m + i)(m - i)$, and so

$$p \mid (m + i)(m - i) \text{ in } \mathbb{Z}[i].$$

If $p \mid (m \pm i)$ in $\mathbb{Z}[i]$, then there are integers u and v with $m \pm i = p(u + iv)$. Comparing the imaginary parts gives $pv = 1$, a contradiction. We conclude that p does not satisfy the analog of Euclid's lemma in Theorem 3.57 (recall that $\mathbb{Z}[i]$ is a PID); it follows from Exercise 3.62 on page 158 that p is not an irreducible element in $\mathbb{Z}[i]$. Hence, there is a factorization

$$p = \alpha\beta \text{ in } \mathbb{Z}[i]$$

in which neither $\alpha = a + ib$ nor $\beta = c + id$ is a unit. Therefore, taking norms gives an equation in $\mathbb{Z}$:

$$\begin{aligned} p^2 &= \partial(p) \\ &= \partial(\alpha\beta) \\ &= \partial(\alpha)\partial(\beta) \\ &= (a^2 + b^2)(c^2 + d^2). \end{aligned}$$

By Proposition 3.64, the only units in $\mathbb{Z}[i]$ are ± 1 and $\pm i$, so that any nonzero Gaussian integer that is not a unit has norm > 1; therefore, $a^2 + b^2 \neq 1$ and $c^2 + d^2 \neq 1$. Euclid's lemma now gives $p \mid (a^2 + b^2)$ or $p \mid (c^2 + d^2)$; the fundamental theorem of arithmetic gives $p = a^2 + b^2$ (and $p = c^2 + d^2$), as desired. •

We are going to determine all the irreducible elements in $\mathbb{Z}[i]$, but we first prove a lemma.

Lemma 3.67. *If $\alpha \in \mathbb{Z}[i]$ is irreducible, then there is a unique prime number p with $\alpha \mid p$ in $\mathbb{Z}[i]$.*

Proof. Note that if $\alpha \in \mathbb{Z}[i]$, then $\overline{\alpha} \in \mathbb{Z}[i]$; since $\partial(\alpha) = \alpha\overline{\alpha}$, we have $\alpha \mid \partial(\alpha)$. Now $\partial(\alpha) = p_1 \cdots p_n$, where the p_i are prime numbers. As $\mathbb{Z}[i]$ is a PID, Exercise 3.62 on page 158 gives $\alpha \mid p_i$ for some i (for α is irreducible). If $\alpha \mid q$ for some prime $q \neq p_i$, then $\alpha \mid (q, p_i) = 1$, forcing α to be a unit. This contradiction shows that p_i is the unique prime number divisible by α. •

Proposition 3.68. *Let $\alpha = a + bi \in \mathbb{Z}[i]$ be neither 0 nor a unit. Then α is irreducible if and only if*

(i) *α is an associate of a prime p in $\mathbb{Z}$ of the form $p = 4m + 3$; or*

(ii) *α is an associate of $1 + i$ or its conjugate $1 - i$; or*

(iii) *$\partial(\alpha) = a^2 + b^2$ is a prime in $\mathbb{Z}$ of the form $4m + 1$.*

Proof. By Lemma 3.67, there is a unique prime number p divisible by α in $\mathbb{Z}[i]$. Since $\alpha \mid p$, we have $\partial(\alpha) \mid \partial(p) = p^2$ in $\mathbb{Z}$, so that $\partial(\alpha) = p$ or $\partial(\alpha) = p^2$; that is,

$$a^2 + b^2 = p \qquad \text{or} \qquad a^2 + b^2 = p^2,$$

Looking at p mod 4, we see that there are three possibilities (for $p \equiv 0 \bmod 4$ cannot occur).

(i) $p \equiv 3 \bmod 4$.

In this case, $a^2 + b^2 = p$ cannot occur, by (the easy direction of) Theorem 3.66, so that $\partial(\alpha) = a^2 + b^2 = p^2$. Now p is divisible by α, so there is β with $\alpha\beta = p$. Hence, $\partial(\alpha)\partial(\beta) = \partial(p)$. Since $p \in \mathbb{Z}$, we have $\partial(p) = p^2$, so that $p^2\partial(\beta) = p^2$. Thus, $\partial(\beta) = 1$, β is a unit, by Proposition 3.64(i), and p is irreducible in $\mathbb{Z}[i]$.

(ii) $p \equiv 2 \bmod 4$.

In this case, $p = 2$, and so $a^2 + b^2 = 2$ or $a^2 + b^2 = 4$. The latter case cannot occur (because a and b are integers), and the first case gives $\alpha = 1 \pm i$ (up to multiplication by units). The reader should check that both $1 + i$ and $1 - i$ are, indeed, irreducible elements.

(iii) $p \equiv 1 \bmod 4$.

If $\partial(\alpha)$ is a prime p (with $p \equiv 1 \bmod 4$), then α is irreducible, by Proposition 3.64(ii). Conversely, suppose α is irreducible. As $\partial(\alpha) = p$ or $\partial(\alpha) = p^2$, it suffices to eliminate the latter possibility. Since $\alpha \mid p$, we have $p = \alpha\beta$ for some $\beta \in \mathbb{Z}[i]$; hence, as in case (i), $\partial(\alpha) = p^2$ implies that β is a unit. Now $\alpha\overline{\alpha} = p^2 = (\alpha\beta)^2$, so that $\overline{\alpha} = \alpha\beta^2$. But $\beta^2 = \pm 1$, by Proposition 3.64(iii), contradicting $\overline{\alpha} \neq \pm\alpha$. Therefore, $\partial(\alpha) = p$. •

For example, 3 is an irreducible element of the first type, and $2 + i$ is an irreducible element of the third type. We should remember that there are interesting connections between prime numbers and irreducible Gaussian integers, that knowing the Gaussian units is valuable, and that the norm is a useful tool in proving results. The ring of Gaussian integers is an instance of a ring of algebraic integers, and these comments remain true for these rings as well.

EXERCISES

Definition. Let k be a field. A ***common divisor*** of $a_1(x), a_2(x), \ldots, a_n(x)$ in $k[x]$ is a polynomial $c(x) \in k[x]$ with $c(x) \mid a_i(x)$ for all i; the ***greatest common divisor*** is the monic common divisor of largest degree. We write $c(x) = (a_1, a_2, \ldots, a_n)$.

3.57 Let k be a field, and let polynomials $a_1(x), a_2(x), \ldots, a_n(x)$ in $k[x]$ be given.

(i) Show that the greatest common divisor $d(x)$ of these polynomials has the form $\sum t_i(x)a_i(x)$, where $t_i(x) \in k[x]$ for $1 \leq i \leq n$.

Hint. Example 3.49.

(ii) Prove that $c(x) \mid d(x)$ for every monic common divisor $c(x)$ of the $a_i(x)$.

3.58 (i) Show that $x, y \in k[x, y]$ are relatively prime, but that 1 is not a linear combination of them [i.e., there do not exist $s(x, y), t(x, y) \in k[x, y]$ with $1 = xs(x, y) + yt(x, y)$].

Hint. Use a degree argument.

(ii) Show that 2 and x are relatively prime in $\mathbb{Z}[x]$, but that 1 is not a linear combination of them; that is, there do not exist $s(x), t(x) \in \mathbb{Z}[x]$ with $1 = 2s(x) + xt(x)$.

3.59 A student claims that $x - 1$ is not irreducible because $x - 1 = (\sqrt{x} + 1)(\sqrt{x} - 1)$ is a factorization. Explain the error of his ways.
Hint. Show that $\sqrt{x} + 1$ is not a polynomial.

3.60 Prove that there are domains R containing a pair of elements having no gcd. (See the definition on page 147.)
Hint. Let k be a field and let R be the subring of $k[x]$ consisting of all polynomials having no linear term; that is, $f(x) \in R$ if and only if

$$f(x) = s_0 + s_2x^2 + s_3x^3 + \cdots .$$

Show that x^5 and x^6 have no gcd in R.

3.61 Prove that $R = \mathbb{Z}[\sqrt{2}] = \{a + b\sqrt{2} : a, b \in \mathbb{Z}\}$ is a euclidean ring with $\partial(a + b\sqrt{2}) = |a^2 - 2b^2|$.

3.62 If R is a euclidean ring and $\pi \in R$ is irreducible, prove that $\pi \mid \alpha\beta$ implies $\pi \mid \alpha$ or $\pi \mid \beta$.

3.63 Let ∂ be the degree function of a euclidean ring R. If $m, n \in \mathbb{N}$ and $m \geq 1$, prove that ∂' is also a degree function on R, where

$$\partial'(x) = m\partial(x) + n$$

for all $x \in R$. Conclude that a euclidean ring may have no elements of degree 0 or degree 1.

3.64 Let R be a euclidean ring with degree function ∂.

(i) Prove that $\partial(1) \leq \partial(a)$ for all nonzero $a \in R$.

(ii) Prove that a nonzero $u \in R$ is a unit if and only if $\partial(u) = \partial(1)$.

Hint. A proof can be generalized from the special case of polynomials.

3.65 Let R be a euclidean ring, and assume that $b \in R$ is neither zero nor a unit. Prove, for every $i \geq 0$, that $\partial(b^i) < \partial(b^{i+1})$.
Hint. There are $q, r \in R$ with $b^i = qb^{i+1} + r$.

3.66 If p is a prime and $p \equiv 3 \bmod 4$, prove that one of the congruences $a^2 \equiv 2 \bmod p$ or $a^2 \equiv -2 \bmod p$ is solvable.
Hint. Show that $\mathbb{I}_p^\times \cong \langle -1 \rangle \times H$, where H is a group of odd order m, say, and observe that either 2 or -2 lies in H because

$$\mathbb{I}_2 \times \mathbb{I}_m = (\{1\} \times H) \cup (\{-1\} \times H) .$$

Finally, use Exercise 2.54 on page 81.

3.7 LINEAR ALGEBRA

We interrupt the exposition to discuss some linear algebra, for it is a necessary tool in further investigation of commutative rings.

Vector Spaces

Linear algebra is the study of vector spaces and their homomorphisms, with applications to systems of linear equations. From now on, we are going to assume that most readers have had some course involving matrices, perhaps only with real entries or with complex entries. Such courses often deal mainly with computational aspects of the subject, such as Gaussian elimination, and finding inverses, determinants, eigenvalues, and characteristic polynomials of matrices, but here we do not emphasize this important aspect of linear algebra. Instead, we discuss more theoretical properties of vector spaces (with scalars in any field) and linear transformations (which are homomorphisms between vector spaces).

Dimension is a rather subtle idea. We think of a curve in the plane, that is, the image of a continuous function $f : \mathbb{R} \to \mathbb{R}^2$, as a one-dimensional subset of a two-dimensional ambient space. Imagine the confusion at the end of the nineteenth century when a "space-filling curve" was discovered: There exists a continuous function $f : \mathbb{R} \to \mathbb{R}^2$ with image the whole plane! We are going to describe a way of defining dimension that works for analogs of euclidean space, called vector spaces (there are topological ways of defining dimension of more general spaces).

Definition. If k is a field, then a ***vector space over k*** is an (additive) abelian group V equipped with a ***scalar multiplication***; that is, there is a function $k \times V \to V$, denoted by $(a, v) \mapsto av$, such that, for all $a, b, 1 \in k$ and all $u, v \in V$,

(i) $a(u + v) = au + av$;

(ii) $(a + b)v = av + bv$;

(iii) $(ab)v = a(bv)$;

(iv) $1v = v$.

The elements of V are called ***vectors*** and the elements of k are called ***scalars***.[13]

Example 3.69.
(i) Euclidean space $V = \mathbb{R}^n$ is a vector space over $\mathbb{R}$. Vectors are n-tuples $(a_1, \ldots, a_n)$, where $a_i \in \mathbb{R}$ for all i. Picture a vector v as an arrow from the origin to the point having coordinates $(a_1, \ldots, a_n)$. Addition is given by

$$(a_1, \ldots, a_n) + (b_1, \ldots, b_n) = (a_1 + b_1, \ldots, a_n + b_n);$$

geometrically, the sum of two vectors is described by the *parallelogram law*.

Scalar multiplication is given by

$$av = a(a_1, \ldots, a_n) = (aa_1, \ldots, aa_n).$$

[13]The word *vector* comes from the Latin word meaning "to carry"; vectors in euclidean space carry the data of length and direction. The word *scalar* comes from regarding $v \mapsto av$ as a change of scale. The terms *scale* and *scalar* come from the Latin word meaning "ladder," for the rungs of a ladder are evenly spaced.

Scalar multiplication $v \mapsto av$ "stretches" v by a factor $|a|$, reversing its direction when a is negative (we put quotes around *stretches* because av is shorter than v when $|a| < 1$).

(ii) The example in part (i) can be generalized. If k is any field, define $V = k^n$, the set of all n-tuples $v = (a_1, \ldots, a_n)$, where $a_i \in k$ for all i. Addition is given by

$$(a_1, \ldots, a_n) + (b_1, \ldots, b_n) = (a_1 + b_1, \ldots, a_n + b_n),$$

and scalar multiplication is given by

$$av = a(a_1, \ldots, a_n) = (aa_1, \ldots, aa_n).$$

(iii) If R is a commutative ring and k is a subring that is a field, then R is a vector space over k. Regard the elements of R as vectors and the elements of k as scalars; define scalar multiplication av, where $a \in k$ and $v \in R$, to be the given product of two elements in R. Notice that the axioms in the definition of vector space are just particular cases of some of the axioms holding in the commutative ring R.

For example, if k is a field, then the polynomial ring $R = k[x]$ is a vector space over k. Vectors are polynomials $f(x)$, scalars are elements $a \in k$, and scalar multiplication gives the polynomial $af(x)$; that is, if

$$f(x) = b_n x^n + \cdots + b_1 x + b_0,$$

then

$$af(x) = ab_n x^n + \cdots + ab_1 x + ab_0.$$

In particular, if a field k is a subfield of a larger field E, then E is a vector space over k. ◀

A *subspace* of a vector space V is a subset of V that is a vector space under the addition and scalar multiplication in V.

Definition. If V is a vector space over a field k, then a ***subspace*** of V is a subset U of V such that

(i) $0 \in U$;

(ii) $u, u' \in U$ imply $u + u' \in U$;

(iii) $u \in U$ and $a \in k$ imply $au \in U$.

Example 3.70.
(i) The extreme cases $U = V$ and $U = \{0\}$ (where $\{0\}$ denotes the subset consisting of the zero vector alone) are always subspaces of a vector space. A subspace $U \subseteq V$ with $U \neq V$ is called a ***proper subspace*** of V; we may write $U \subsetneq V$ to denote U being a proper subspace of V.

(ii) If $v = (a_1, \ldots, a_n)$ is a nonzero vector in $\mathbb{R}^n$, then the line through the origin

$$\ell = \{av : a \in \mathbb{R}\}$$

is a subspace of $\mathbb{R}^n$.

Similarly, a plane through the origin consists of all vectors of the form $av_1 + bv_2$, where v_1, v_2 is a fixed pair of noncollinear vectors, and a, b vary over $\mathbb{R}$. It is easy to check that planes through the origin are subspaces of $\mathbb{R}^n$.

(iii) If $m \le n$ and $\mathbb{R}^m$ is regarded as the set of all those vectors in $\mathbb{R}^n$ whose last $n - m$ coordinates are 0, then $\mathbb{R}^m$ is a subspace of $\mathbb{R}^n$. For example, we may regard the plane $\mathbb{R}^2$ as all points $(x, y, 0)$ in $\mathbb{R}^3$.

(iv) If k is a field, then a ***homogeneous linear system over*** k of m equations in n unknowns is a set of equations

$$\begin{aligned} a_{11}x_1 + \cdots + a_{1n}x_n &= 0 \\ a_{21}x_1 + \cdots + a_{2n}x_n &= 0 \\ \vdots \qquad\qquad & \vdots \\ a_{m1}x_1 + \cdots + a_{mn}x_n &= 0, \end{aligned}$$

where $a_{ji} \in k$. A ***solution*** of this system is a vector $(c_1, \ldots, c_n) \in k^n$, where $\sum_i a_{ji}c_i = 0$ for all j; a solution $(c_1, \ldots, c_n)$ is ***nontrivial*** if some $c_i \ne 0$. The set of all solutions forms a subspace of k^n, called the ***solution space*** (or *nullspace*) of the system.

In particular, we can solve systems of linear equations over $\mathbb{I}_p$, where p is a prime. This says that we can treat a system of congruences mod p just as one treats an ordinary system of equations.

For example, the system of congruences

$$\begin{aligned} 3x - 2y + z &\equiv 1 \bmod 7 \\ x + y - 2z &\equiv 0 \bmod 7 \\ -x + 2y + z &\equiv 4 \bmod 7 \end{aligned}$$

can be regarded as a system of equations over the field $\mathbb{I}_7$. This system can be solved just as in high school, for inverses mod 7 are now known: $[2][4] = [1]$; $[3][5] = [1]$; $[6][6] = [1]$. The solution is

$$(x, y, z) = ([5], [4], [1]). \quad ◀$$

Definition. A ***list*** in a vector space V is an ordered set $v_1, \ldots, v_n$ of vectors in V.

More precisely, we are saying that there is some $n \ge 1$ and some function

$$\varphi \colon \{1, 2, \ldots, n\} \to V,$$

with $\varphi(i) = v_i$ for all i. Thus, $X = \operatorname{im} \varphi$; note that X is ordered in the sense that there is a first vector v_1, a second vector v_2, and so forth. A vector may appear several times on a list; that is, φ need not be injective.

Definition. Let V be a vector space over a field k. A ***k-linear combination*** of a list $v_1, \ldots, v_n$ in V is a vector v of the form

$$v = a_1 v_1 + \cdots + a_n v_n,$$

where $a_i \in k$ for all i.

Definition. If $X = v_1, \ldots, v_m$ is a list in a vector space V, then

$$\langle v_1, \ldots, v_m \rangle,$$

the set of all the k-linear combinations of $v_1, \ldots, v_m$, is called the ***subspace spanned by X***. We also say that $v_1, \ldots, v_m$ ***spans*** $\langle v_1, \ldots, v_m \rangle$.

Lemma 3.71. *Let V be a vector space over a field k.*

(i) *Every intersection of subspaces of V is itself a subspace.*

(ii) *If $X = v_1, \ldots, v_m$ is a list in V, then the intersection of all the subspaces of V containing X is $\langle v_1, \ldots, v_m \rangle$, the subspace spanned by $v_1, \ldots, v_m$, and so $\langle v_1, \ldots, v_m \rangle$ is the* ***smallest*** subspace *of V containing X.*

Sketch of Proof. Part (i) is routine. Let $X = \{v_1, \ldots, v_m\}$, and let $\mathcal{S}$ denote the family of all the subspaces of V containing X; we claim that

$$\bigcap_{S \in \mathcal{S}} S = \langle v_1, \ldots, v_m \rangle.$$

The inclusion $\subseteq$ is clear, because $\langle v_1, \ldots, v_m \rangle \in \mathcal{S}$. For the reverse inclusion, note that if $S \in \mathcal{S}$, then S contains $v_1, \ldots, v_m$, and so it contains the set of all linear combination of $v_1, \ldots, v_m$, namely, $\langle v_1, \ldots, v_m \rangle$. •

It follows from the second part of the lemma that the subspace spanned by a list $X = v_1, \ldots, v_m$ does not depend on the ordering of the vectors, but only on the set of vectors themselves. Were all terminology in algebra consistent, we would call $\langle v_1, \ldots, v_m \rangle$ the subspace *generated by X*. The reason for the different terms is that the theories of groups, rings, and vector spaces developed independently of each other.

If $X = \varnothing$, then $\langle X \rangle = \bigcap_{S \in \mathcal{S}} S$, where $\mathcal{S}$ is the family of all the subspaces of V containing X. As every subspace contains $X = \varnothing$, $\{0\}$ itself is one of the subspaces occurring in the intersection of all the subspaces of V, and so $\langle \varnothing \rangle = \bigcap_{S \subseteq V} S = \{0\}$.

Example 3.72.

(i) Let $V = \mathbb{R}^2$, let $e_1 = (1, 0)$, and let $e_2 = (0, 1)$. Now $V = \langle e_1, e_2 \rangle$, for if $v = (a, b) \in V$, then

$$\begin{aligned} v &= (a, 0) + (0, b) \\ &= a(1, 0) + b(0, 1) \\ &= ae_1 + be_2 \in \langle e_1, e_2 \rangle. \end{aligned}$$

(ii) If k is a field and $V = k^n$, define e_i as the n-tuple having 1 in the ith coordinate and 0's elsewhere. The reader may adapt the argument in part (i) to show that $e_1, \ldots, e_n$ spans k^n.

(iii) A vector space V need not be spanned by a finite list. For example, let $V = k[x]$, and suppose that $X = f_1(x), \ldots, f_m(x)$ is a finite list in V. If d is the largest degree of any of the $f_i(x)$, then every (nonzero) k-linear combination of $f_1(x), \ldots, f_m(x)$ has degree at most d. Thus, x^{d+1} is not a k-linear combination of vectors in X, and so X does not span $k[x]$. ◂

The following definition makes sense even though we have not yet defined *dimension*.

Definition. A vector space V is called ***finite-dimensional*** if it is spanned by a finite list; otherwise, V is called ***infinite-dimensional***.

Example 3.72(ii) shows that k^n is finite-dimensional, while part (iii) of this Example shows that $k[x]$ is infinite-dimensional. By Example 3.69(iii), both $\mathbb{R}$ and $\mathbb{C}$ are vector spaces over $\mathbb{Q}$, and they are both infinite-dimensional.

Notation. If $v_1, \ldots, v_m$ is a list, then $v_1, \ldots, \widehat{v_i} \ldots, v_m$ is the shorter list with v_i deleted.

Proposition 3.73. *If V is a vector space, then the following conditions on a list $X = v_1, \ldots, v_m$ spanning V are equivalent:*

(i) *X is not a shortest spanning list;*

(ii) *some v_i is in the subspace spanned by the others; that is,*

$$v_i \in \langle v_1, \ldots, \widehat{v_i}, \ldots, v_m \rangle ;$$

(iii) *there are scalars $a_1, \ldots, a_m$, not all zero, with*

$$\sum_{\ell=1}^{m} a_\ell v_\ell = 0.$$

Sketch of Proof. (i) ⇒ (ii). If X is not a shortest spanning list, then one of the vectors in X can be thrown out, and the shorter list still spans.

(ii) ⇒ (iii). If $v_i = \sum_{j \neq i} c_j v_j$, then define $a_i = -1 \neq 0$ and $a_j = c_j$ for all $j \neq i$.

(iii) ⇒ (i). The given equation implies that one of the vectors, say, v_i, is a linear combination of the others. Deleting v_i gives a shorter list, which still spans: If $v \in V$ is a linear combination of all the v_j (including v_i), just substitute the expression for v_i as a linear combination of the other v_j and collect terms. •

Definition. A list $X = v_1, \ldots, v_m$ in a vector space V is ***linearly dependent*** if there are scalars $a_1, \ldots, a_m$, not all zero, with $\sum_{\ell=1}^{m} a_\ell v_\ell = 0$; otherwise, X is called ***linearly independent***.

The empty set $\varnothing$ is defined to be linearly independent (we may interpret $\varnothing$ as a list of length 0).

Example 3.74.

(i) Any list $X = v_1, \ldots, v_m$ containing the zero vector is linearly dependent.

(ii) A list v_1 of length 1 is linearly dependent if and only if $v_1 = 0$; hence, a list v_1 of length 1 is linearly independent if and only if $v_1 \neq 0$.

(iii) A list v_1, v_2 is linearly dependent if and only if one of the vectors is a scalar multiple of the other.

(iv) If there is a repetition in the list $v_1, \ldots, v_m$ (that is, if $v_i = v_j$ for some $i \neq j$), then $v_1, \ldots, v_m$ is linearly dependent: Define $c_i = 1$, $c_j = -1$, and all other $c = 0$. Therefore, if $v_1, \ldots, v_m$ is linearly independent, then all the vectors v_i are distinct. ◄

The contrapositive of Proposition 3.73 is worth stating.

Corollary 3.75. *If $X = v_1, \ldots, v_m$ is a list spanning a vector space V, then X is a shortest spanning list if and only if X is linearly independent.*

Linear independence has been defined indirectly, as not being linearly dependent. Because of the importance of linear independence, let us define it directly. A list $X = v_1, \ldots, v_m$ is ***linearly independent*** if, whenever a k-linear combination $\sum_{\ell=1}^{m} a_\ell v_\ell = 0$, then every $a_i = 0$. It follows that every sublist of a linearly independent list is itself linearly independent (this is one reason for decreeing that $\varnothing$ be linearly independent).

We have arrived at the notion we have been seeking.

Definition. A ***basis*** of a vector space V is a linearly independent list that spans V.

Thus, bases are shortest spanning lists. Of course, all the vectors in a linearly independent list $v_1, \ldots, v_n$ are distinct, by Example 3.74(iv).

Example 3.76.

In Example 3.72(ii), we saw that $X = e_1, \ldots, e_n$ spans k^n, where e_i is the n-tuple having 1 in the ith coordinate and 0's elsewhere. We can easily prove that X is linearly independent, and hence it is a basis; it is called the ***standard basis*** of k^n. ◄

Proposition 3.77. *Let $X = v_1, \ldots, v_n$ be a list in a vector space V over a field k. Then X is a basis if and only if each vector in V has a unique expression as a k-linear combination of vectors in X.*

Sketch of Proof. If a vector $v = \sum a_i v_i = \sum b_i v_i$, then $\sum(a_i - b_i)v_i = 0$, and so independence gives $a_i = b_i$ for all i; that is, the expression is unique.

Conversely, existence of an expression shows that the list of v_i spans. Moreover, if $0 = \sum c_i v_i$ with not all $c_i = 0$, then the vector 0 does not have a unique expression as a linear combination of the v_i. •

Definition. If $X = v_1, \ldots, v_n$ is a basis of a vector space V and if $v \in V$, then there are unique scalars $a_1, \ldots, a_n$ with $v = \sum_{i=1}^{n} a_i v_i$. The n-tuple $(a_1, \ldots, a_n)$ is called the ***coordinate set*** of a vector $v \in V$ relative to the basis X.

Observe that if $v_1, \ldots, v_n$ is the standard basis of $V = k^n$, then this coordinate set coincides with the usual coordinate set.

If $v_1, \ldots, v_n$ is a basis of a vector space V over a field k, then each vector $v \in V$ has a unique expression

$$v = a_1 v_1 + a_2 v_2 + \cdots + a_n v_n,$$

where $a_i \in k$ for all i. Since there is a first vector v_1, a second vector v_2, and so forth, the coefficients in this k-linear combination determine a unique n-tuple $(a_1, a_2, \ldots, a_n)$. Were a basis merely a subset of V and not a list (i.e., an ordered subset), then there would be $n!$ coordinate sets for every vector.

We are going to define the *dimension* of a vector space V to be the number of vectors in a basis. Two questions arise at once.

(i) Does every vector space have a basis?

(ii) Do all bases of a vector space have the same number of elements?

The first question is easy to answer; the second needs some thought.

Theorem 3.78. *Every finite-dimensional vector space V has a basis.*

Sketch of Proof. A finite spanning list X exists, since V is finite-dimensional. If it is linearly independent, it is a basis; if not, X can be shortened to a spanning sublist X', by Proposition 3.73. If X' is linearly independent, it is a basis; if not, X' can be shortened to a spanning sublist X''. Eventually, we arrive at a shortest spanning sublist, which is independent and hence is a basis. •

The definitions of spanning and linear independence can be extended to infinite lists in a vector space, and we can then prove that infinite-dimensional vector spaces also have bases (see Theorem 6.48). For example, it turns out that a basis of $k[x]$ is $1, x, x^2, \ldots, x^n, \ldots$.

We can now prove invariance of dimension, one of the most important results about vector spaces.

Lemma 3.79. *Let $u_1, \ldots, u_n$ be elements in a vector space V, and let $v_1, \ldots, v_m \in \langle u_1, \ldots, u_n \rangle$. If $m > n$, then $v_1, \ldots, v_m$ is a linearly dependent list.*

Proof. The proof is by induction on $n \geq 1$.
Base Step. If $n = 1$, then there are at least two vectors v_1, v_2 and $v_1 = a_1 u_1$ and $v_2 = a_2 u_2$. If $u_1 = 0$, then $v_1 = 0$ and the list of v's is linearly dependent. Suppose $u_1 \neq 0$. We may assume that $v_1 \neq 0$, or we are done; hence, $a_1 \neq 0$. Therefore, v_1, v_2 is linearly dependent, for $v_2 - a_2 a_1^{-1} v_1 = 0$, and hence the larger list $v_1, \ldots, v_m$ is linearly dependent.

Inductive Step. There are equations, for $i = 1, \ldots, m$,

$$v_i = a_{i1} u_1 + \cdots + a_{in} u_n.$$

We may assume that some $a_{i1} \neq 0$, otherwise $v_1, \ldots, v_m \in \langle u_2, \ldots, u_n \rangle$, and the inductive hypothesis applies. Changing notation if necessary (that is, by re-ordering the v's), we may assume that $a_{11} \neq 0$. For each $i \geq 2$, define

$$v_i' = v_i - a_{i1} a_{11}^{-1} v_1 \in \langle u_2, \ldots, u_n \rangle$$

(writing v_i' as a linear combination of the u's, the coefficient of u_1 is $a_{i1} - (a_{i1} a_{11}^{-1}) a_{11} = 0$). Since $m - 1 > n - 1$, the inductive hypothesis gives scalars $b_2, \ldots, b_m$, not all 0, with

$$b_2 v_2' + \cdots + b_m v_m' = 0.$$

Rewrite this equation using the definition of v_i':

$$\Big(-\sum_{i \geq 2} b_i a_{i1} a_{11}^{-1}\Big) v_1 + b_2 v_2 + \cdots + b_m v_m = 0.$$

Not all the coefficients are 0, and so $v_1, \ldots, v_m$ is linearly dependent. •

The following familiar fact illustrates the intimate relation between linear algebra and systems of linear equations.

Corollary 3.80. *A homogeneous system of linear equations, over a field k, with more unknowns than equations has a nontrivial solution.*

Proof. An n-tuple $(\beta_1, \ldots, \beta_n)$ is a solution of a system

$$\begin{aligned} \alpha_{11} x_1 + \cdots + \alpha_{1n} x_n &= 0 \\ \vdots \qquad \vdots \qquad &\vdots \\ \alpha_{m1} x_1 + \cdots + \alpha_{mn} x_n &= 0 \end{aligned}$$

if $\alpha_{i1} \beta_1 + \cdots + \alpha_{in} \beta_n = 0$ for all i. In other words, if $c_1, \ldots, c_n$ are the columns of the $m \times n$ coefficient matrix $A = [\alpha_{ij}]$, then

$$\beta_1 c_1 + \cdots + \beta_n c_n = 0.$$

Note that $c_i \in k^m$. Now k^m can be spanned by m vectors (the standard basis, for example). Since $n > m$, by hypothesis, Lemma 3.79 shows that the list $c_1, \ldots, c_n$ is linearly dependent; there are scalars $\gamma_1, \ldots, \gamma_n$, not all zero, with $\gamma_1 c_1 + \cdots + \gamma_n c_n = 0$. Therefore, $(\gamma_1, \ldots, \gamma_n)$ is a nontrivial solution of the system. •

Theorem 3.81 (Invariance of Dimension). *If $X = x_1, \ldots, x_n$ and $Y = y_1, \ldots, y_m$ are bases of a vector space V, then $m = n$.*

Proof. If $m \neq n$, then either $n < m$ or $m < n$. In the first case, $y_1, \ldots, y_m \in \langle x_1, \ldots, x_n \rangle$, because X spans V, and Lemma 3.79 gives Y linearly dependent, a contradiction. A similar contradiction arises if $m < n$, and so we must have $m = n$. •

It is now permissible to make the following definition.

Definition. If V is a finite-dimensional vector space over a field k, then its ***dimension***, denoted by $\dim_k(V)$ or $\dim(V)$, is the number of elements in a basis of V.

Example 3.82.
(i) Example 3.76 shows that k^n has dimension n, which agrees with our intuition when $k = \mathbb{R}$. Thus, the plane $\mathbb{R} \times \mathbb{R}$ is two-dimensional!

(ii) If $V = \{0\}$, then $\dim(V) = 0$, for there are no elements in its basis $\varnothing$. (This is a good reason for defining $\varnothing$ to be linearly independent.)

(iii) Let $X = \{x_1, \ldots, x_n\}$ be a finite set. Define

$$k^X = \{\text{functions } f : X \to k\}.$$

Now k^X is a vector space if we define addition $f + f'$ to be

$$f + f' : x \mapsto f(x) + f'(x)$$

and scalar multiplication af, for $a \in k$ and $f : X \to k$, by

$$af : x \mapsto af(x).$$

It is easy to check that the set of n functions of the form f_x, where $x \in X$, defined by

$$f_x(y) = \begin{cases} 1 & \text{if } y = x; \\ 0 & \text{if } y \neq x, \end{cases}$$

form a basis, and so $\dim(k^X) = n = |X|$.

The reader should note that this is not a new example: An n-tuple $(a_1, \ldots, a_n)$ is really a function $f : \{1, \ldots, n\} \to k$ with $f(i) = a_i$ for all i. Thus, the functions f_x comprise the standard basis. ◂

Here is a second proof of invariance of dimension; it will be used, in Chapter 6, to generalize the notion of dimension to the notion of *transcendence degree*. We begin with a modification of the proof of Proposition 3.73.

Lemma 3.83. *If $X = v_1, \ldots, v_n$ is a linearly dependent list of vectors in a vector space V, then there exists v_r with $r \geq 1$ with $v_r \in \langle v_1, v_2, \ldots, v_{r-1} \rangle$ [when $r = 1$, we interpret $\langle v_1, \ldots, v_{r-1} \rangle$ to mean $\{0\}$].*

Remark. Let us compare Proposition 3.73 with this one. The earlier result says that if v_1, v_2, v_3 is linearly dependent, then either $v_1 \in \langle v_2, v_3 \rangle$, $v_2 \in \langle v_1, v_3 \rangle$, or $v_3 \in \langle v_1, v_2 \rangle$. This lemma says that either $v_1 \in \{0\}$, $v_2 \in \langle v_1 \rangle$, or $v_3 \in \langle v_1, v_2 \rangle$. ◀

Proof. Let r be the largest integer for which $v_1, \ldots, v_{r-1}$ is linearly independent. If $v_1 = 0$, then $v_1 \in \{0\}$, and we are done. If $v_1 \neq 0$, then $r \geq 2$; since $v_1, v_2, \ldots, v_n$ is linearly dependent, we have $r - 1 < n$. As $r - 1$ is largest, the list $v_1, v_2, \ldots, v_r$ is linearly dependent. There are thus scalars $a_1, \ldots, a_r$, not all zero, with $a_1 v_1 + \cdots + a_r v_r = 0$. In this expression, we must have $a_r \neq 0$, for otherwise $v_1, \ldots, v_{r-1}$ would be linearly dependent. Therefore,

$$v_r = \sum_{i=1}^{r-1} (-a_r^{-1}) a_i v_i \in \langle v_1, \ldots, v_{r-1} \rangle . \quad \bullet$$

Lemma 3.84 (Exchange Lemma). *If $X = x_1, \ldots, x_m$ is a basis of a vector space V and $y_1, \ldots, y_n$ is a linearly independent subset of V, then $n \leq m$.*

Proof. We begin by showing that one of the x's in X can be replaced by y_n so that the new list still spans V. Now $y_n \in \langle X \rangle$, since X spans V, so that the list

$$y_n, x_1, \ldots, x_m$$

is linearly dependent, by Proposition 3.73. Since the list $y_1, \ldots, y_n$ is linearly independent, $y_n \notin \{0\}$. By Lemma 3.83, there is some i with $x_i = a y_n + \sum_{j<i} a_j x_j$. Throwing out x_i and replacing it by y_n gives a spanning list

$$X' = y_n, x_1, \ldots, \widehat{x_i}, \ldots, x_m :$$

If $v = \sum_{j=1}^m b_j x_j$, then (as in the proof of Proposition 3.73), replace x_i by its expression as a k-linear combination of the other x's and y_n, and then collect terms.

Now repeat this argument for the spanning list $y_{n-1}, y_n, x_1, \ldots, \widehat{x_i}, \ldots, x_m$. The options offered by Lemma 3.83 for this linearly dependent list are $y_n \in \langle y_{n-1} \rangle$, $x_1 \in \langle y_{n-1}, y_n \rangle$, $x_2 \in \langle y_{n-1}, y_n, x_1 \rangle$, and so forth. Since Y is linearly independent, so is its sublist y_{n-1}, y_n, and the first option $y_n \in \langle y_{n-1} \rangle$ is not feasible. It follows that the disposable vector (provided by Lemma 3.83) must be one of the remaining x's, say x_ℓ. After throwing out x_ℓ, we have a new spanning list X''. Repeat this construction of spanning lists; each time a new y is adjoined as the first vector, an x is thrown out, for the option $y_i \in \langle y_{i+1}, \ldots, y_n \rangle$ is not feasible. If $n > m$, that is, if there are more y's than x's, then this procedure ends with a spanning list consisting of m y's (one for each of the m x's thrown out) and no x's. Thus a proper sublist of $Y = y_1, \ldots, y_n$ spans V, and this contradicts the linear independence of Y. Therefore, $n \leq m$. $\bullet$

Theorem 3.85 (Invariance of Dimension). *If $X = x_1, \ldots, x_m$ and $Y = y_1, \ldots, y_n$ are bases of a vector space V, then $m = n$.*

Proof. By Lemma 3.84, viewing X as a basis with m elements and and Y as a linearly independent list with n elements gives the inequality $n \leq m$; viewing Y a basis and X as a linearly independent list gives the reverse inequality $m \leq n$. Therefore, $m = n$, as desired. •

Definition. A ***longest*** (or a ***maximal***) linearly independent list $u_1, \ldots, u_m$ is a linearly independent list for which there is no vector $v \in V$ such that $u_1, \ldots, u_m, v$ is linearly independent.

Lemma 3.86. *If V is a finite-dimensional vector space, then a longest linearly independent list $v_1, \ldots, v_n$ is a basis of V.*

Sketch of Proof. If the list is not a basis, then it does not span: There is $w \in V$ with $w \notin \langle v_1, \ldots, v_n \rangle$. But the longer list with w adjoined is linearly independent, by Proposition 3.73. •

It is not obvious that there are any longest linearly independent lists; that they do exist follows from the next result, which is quite useful in its own right.

Proposition 3.87. *Let $Z = u_1, \ldots, u_m$ be a linearly independent list in an n-dimensional vector space V. Then Z can be extended to a basis; i.e., there are vectors $v_{m+1}, \ldots, v_n$ so that $u_1, \ldots, u_m, v_{m+1}, \ldots, v_n$ is a basis of V.*

Sketch of Proof. If the linearly independent list Z does not span V, there is $w_1 \in V$ with $w_1 \notin \langle Z \rangle$, and the longer list Z, w_1 is linearly independent, by Proposition 3.73. If Z, w_1 does not span V, there is $w_2 \in V$ with $w_2 \notin \langle Z, w_1 \rangle$. Since $\dim(V) = n$, the length of these lists can never exceed n. Otherwise, compare a linearly independent list with $n + 1$ elements with a basis, and reach a contradiction using the exchange lemma, Lemma 3.84. •

Corollary 3.88. *If $\dim(V) = n$, then any list of $n + 1$ or more vectors is linearly dependent.*

Sketch of Proof. Otherwise, such a list could be extended to a basis having too many elements. •

Corollary 3.89. *Let V be a vector space with $\dim(V) = n$.*

(i) *A list of n vectors that spans V must be linearly independent.*

(ii) *Any linearly independent list of n vectors must span V.*

Sketch of Proof. (i) Were it linearly dependent, then the list could be shortened to give a basis, and this basis is too small.

(ii) If the list does not span, the it could be lengthened to give a basis, and this basis is too large. •

Corollary 3.90. *Let U be a subspace of a vector space V of dimension n.*

(i) *U is finite-dimensional and* $\dim(U) \le \dim(V)$.

(ii) *If* $\dim(U) = \dim(V)$, *then* $U = V$.

Sketch of Proof. (i) Take $u_1 \in U$. If $U = \langle u_1 \rangle$, then U is finite-dimensional. Otherwise, there is $u_2 \notin \langle u_1 \rangle$. By Proposition 3.73, u_1, u_2 is linearly independent. If $U = \langle u_1, u_2 \rangle$, we are done. This process cannot be repeated $n+1$ times, for then $u_1, \ldots, u_{n+1}$ would be a linearly independent list in $U \subseteq V$, contradicting Corollary 3.88.

A basis of U is linearly independent, and so it can be extended to a basis of V.

(ii) If $\dim(U) = \dim(V)$, then a basis of U is already a basis of V (otherwise it could be extended to a basis of V that would be too large). •

EXERCISES

3.67 If the only subspaces of a vector space V are $\{0\}$ and V itself, prove that $\dim(V) \le 1$.

3.68 Prove, in the presence of all the other axioms in the definition of vector space, that the commutative law for vector addition is redundant; that is, if V satisfies all the other axioms, then $u + v = v + u$ for all $u, v \in V$.
Hint. If $u, v \in V$, evaluate $-[(-v) + (-u)]$ in two ways.

3.69 If V is a vector space over $\mathbb{I}_2$ and if $v_1 \ne v_2$ are nonzero vectors in V, prove that v_1, v_2 is linearly independent. Is this true for vector spaces over any other field?

3.70 Prove that the columns of an $m \times n$ matrix A over a field k are linearly dependent in k^m if and only if the homogeneous system $Ax = 0$ has a nontrivial solution.

3.71 If U is a subspace of a vector space V over a field k, define a scalar multiplication on the quotient group V/U by

$$\alpha(v + U) = \alpha v + U,$$

where $\alpha \in k$ and $v \in V$. Prove that this is a well-defined function that makes V/U into a vector space over k (V/U is called a ***quotient space***).

3.72 If V is a finite-dimensional vector space and U is a subspace, prove that

$$\dim(U) + \dim(V/U) = \dim(V).$$

Hint. Prove that if $v_1 + U, \ldots, v_r + U$ is a basis of V/U, then the list $v_1, \ldots, v_r$ is linearly independent.

Definition. If U and W are subspaces of a vector space V, define

$$U + W = \{u + w : u \in U \text{ and } w \in W\}.$$

3.73 (i) Prove that $U + W$ is a subspace of V.

(ii) If U and U' are subspaces of a finite-dimensional vector space V, prove that

$$\dim(U) + \dim(U') = \dim(U \cap U') + \dim(U + U').$$

Hint. Take a basis of $U \cap U'$ and extend it to bases of U and of U'.

Definition. If U and W are vector spaces over a field k, then their ***direct sum*** is the set of all ordered pairs,

$$U \oplus W = \{(u, w) : u \in U \text{ and } w \in W\},$$

with addition

$$(u, w) + (u', w') = (u + u', w + w')$$

and scalar multiplication

$$\alpha(u, w) = (\alpha u, \alpha w).$$

3.74 If U and W are finite-dimensional vector spaces over a field k, prove that

$$\dim(U \oplus W) = \dim(U) + \dim(W).$$

Linear Transformations

Homomorphisms between vector spaces are called *linear transformations*.

Definition. If V and W are vector spaces over a field k, then a function $T: V \to W$ is a ***linear transformation*** if, for all vectors $u, v \in V$ and all scalars $a \in k$,

(i) $T(u + v) = T(u) + T(v)$;

(ii) $T(av) = aT(v)$.

We say that a linear transformation T is ***nonsingular*** (or is an ***isomorphism***) if T is a bijection. Two vector spaces V and W over k are ***isomorphic***, denoted by $V \cong W$, if there is a nonsingular linear transformation $T: V \to W$.

If we forget the scalar multiplication, then a vector space is an (additive) abelian group and a linear transformation T is a group homomorphism. It is easy to see that T preserves all k-linear combinations:

$$T(a_1v_1 + \cdots + a_mv_m) = a_1T(v_1) + \cdots + a_mT(v_m).$$

Example 3.91.

(i) The identity function $1_V : V \to V$ on any vector space V is a nonsingular linear transformation.

(ii) If θ is an angle, then rotation about the origin by θ is a linear transformation $R_\theta : \mathbb{R}^2 \to \mathbb{R}^2$. The function R_θ preserves addition because it takes parallelograms to parallelograms, and it preserves scalar multiplication because it preserves the lengths of arrows.

(iii) If V and W are vector spaces over a field k, write $\text{Hom}_k(V, W)$ for the set of all linear transformations $V \to W$. Define *addition* $S + T$ by $v \mapsto S(v) + T(v)$ for all $v \in V$, and define *scalar multiplication* $\alpha T : V \to W$, where $\alpha \in k$, by $v \mapsto \alpha T(v)$ for all $v \in V$. Both $S + T$ and αT are linear transformations, and $\text{Hom}_k(V, W)$ is a vector space over k. ◀

Definition. If V is a vector space over a field k, then the ***general linear group***, denoted by $\text{GL}(V)$, is the set of all nonsingular linear transformations $V \to V$.

A composite ST of linear transformations S and T is again a linear transformation, and ST is nonsingular if both S and T are; moreover, the inverse of a nonsingular linear transformation is again nonsingular. It follows that $\text{GL}(V)$ is a group with composition as operation, for composition of functions is always associative.

We now show how to construct linear transformations $T : V \to W$, where V and W are vector spaces over a field k. The next theorem says that there is a linear transformation that does anything to a basis.

Theorem 3.92. *Let $v_1, \ldots, v_n$ be a basis of a vector space V over a field k. If W is a vector space over k and $u_1, \ldots, u_n$ is a list in W, then there exists a unique linear transformation $T : V \to W$ with $T(v_i) = u_i$ for all i.*

Proof. By Theorem 3.77, each $v \in V$ has a unique expression of the form $v = \sum_i a_i v_i$, and so $T : V \to W$, given by $T(v) = \sum a_i u_i$, is a (well-defined!) function. It is now a routine verification to check that T is a linear transformation.

To prove uniqueness of T, assume that $S : V \to W$ is a linear transformation with

$$S(v_i) = u_i = T(v_i)$$

for all i. If $v \in V$, then $v = \sum a_i v_i$ and

$$\begin{aligned} S(v) &= S\Big(\sum a_i v_i\Big) \\ &= \sum S(a_i v_i) \\ &= \sum a_i S(v_i) \\ &= \sum a_i T(v_i) = T(v). \end{aligned}$$

Since v is arbitrary, $S = T$. •

Corollary 3.93. *If two linear transformations $S, T : V \to W$ agree on a basis, then $S = T$.*

Proof. This follows at once from the uniqueness of the defined linear transformation. •

Linear transformations defined on k^n are easy to describe.

Proposition 3.94. *If $T : k^n \to k^m$ is a linear transformation, then there exists an $m \times n$ matrix A such that*

$$T(y) = Ay$$

for all $y \in k^n$ (here, y is an $n \times 1$ column matrix and Ay is matrix multiplication).

Sketch of Proof. If $e_1, \ldots, e_n$ is the standard basis of k^n and $e'_1, \ldots, e'_m$ is the standard basis of k^m, define $A = [a_{ij}]$ to be the matrix whose jth column is the coordinate set of $T(e_j)$. If $S: k^n \to k^m$ is defined by $S(y) = Ay$, then $S = T$ because both agree on a basis: $T(e_j) = \sum_i a_{ij} e_i = Ae_j$. •

Theorem 3.92 establishes the connection between linear transformations and matrices, and the definition of matrix multiplication arises from applying this construction to the composite of two linear transformations.

Definition. Let $X = v_1, \ldots, v_n$ be a basis of V and let $Y = w_1, \ldots, w_m$ be a basis of W. If $T : V \to W$ is a linear transformation, then the ***matrix of T*** is the $m \times n$ matrix $A = [a_{ij}]$ whose jth column $a_{1j}, a_{2j}, \ldots, a_{mj}$ is the coordinate set of $T(v_j)$ determined by the w's: $T(v_j) = \sum_{i=1}^{m} a_{ij} w_i$. The matrix A does depend on the choice of bases X and Y; we will write

$$A = {}_Y[T]_X$$

when it is necessary to display them.

In case $V = W$, we often let the bases $X = v_1, \ldots, v_n$ and $w_1, \ldots, w_m$ coincide. If $1_V : V \to V$, given by $v \mapsto v$, is the identity linear transformation, then ${}_X[1_V]_X$ is the $n \times n$ ***identity matrix*** I_n (usually, the subscript n is omitted), defined by

$$I = [\delta_{ij}],$$

where δ_{ij} is the Kronecker delta. Thus, I has 1's on the diagonal and 0's elsewhere. On the other hand, if X and Y are different bases, then ${}_Y[1_V]_X$ is not the identity matrix; its columns are the coordinate sets of the x's with respect to the basis Y.

Example 3.95.
Let $T : V \to W$ be a linear transformation, and let $X = v_1, \ldots, v_n$ and $Y = w_1, \ldots, w_m$ be bases of V and W, respectively. The matrix for T is set up from the equation

$$T(v_j) = a_{1j} w_1 + a_{2j} w_2 + \cdots + a_{mj} w_m.$$

Why are the indices reversed? Why not write

$$T(v_j) = a_{j1}w_1 + a_{j2}w_2 + \cdots + a_{jm}w_m?$$

Consider the following example. Let A be an $m \times n$ matrix over a field k. The function $T : k^n \to k^m$, defined by $T(X) = AX$, where X is an $n \times 1$ column vector, is a linear transformation. If $e_1, \ldots, e_n$ and $e'_1, \ldots, e'_m$ are the standard bases of k^n and k^m, respectively, then the definition of matrix multiplication says that $T(e_j) = Ae_j$ is the jth column of A. But

$$Ae_j = a_{1j}e'_1 + a_{2j}e'_2 + \cdots + a_{mj}e'_m.$$

Therefore, the matrix associated to T is the original matrix A.

In Proposition 3.98, we shall prove that matrix multiplication arises from composition of linear transformations. If $T : V \to W$ has matrix A and $S : W \to U$ has matrix B, then the linear transformation $ST : V \to U$ has matrix BA. Had we defined matrices of linear transformations by making coordinate sets rows instead of columns, then the matrix of ST would have been AB. ◂

Example 3.96.
(i) Let $T : \mathbb{R}^2 \to \mathbb{R}^2$ be rotation by 90°. The matrix of T relative to the standard basis $X = (1, 0), (0, 1)$ is

$${}_X[T]_X = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}.$$

However, if $Y = (0, 1), (1, 0)$, then

$${}_Y[T]_Y = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}.$$

(ii) Let k be a field, let $T : V \to V$ be a linear transformation on a two-dimensional vector space, and assume that there is some vector $v \in V$ with $T(v)$ not a scalar multiple of v. The assumption on v says that the list $X = v, T(v)$ is linearly independent, by Example 3.74(iii), and hence it is a basis of V [because $\dim(V) = 2$]. Write $v_1 = v$ and $v_2 = Tv$.

We compute ${}_X[T]_X$.

$$T(v_1) = v_2 \quad \text{and} \quad T(v_2) = av_1 + bv_2$$

for some $a, b \in k$. We conclude that

$${}_X[T]_X = \begin{bmatrix} 0 & a \\ 1 & b \end{bmatrix}. \quad ◂$$

The following proposition is a paraphrase of Theorem 3.92.

Proposition 3.97. *Let V and W be vector spaces over a field k, and let $X = v_1, \ldots, v_n$ and $Y = w_1, \ldots, w_m$ be bases of V and W, respectively. If $\mathrm{Hom}_k(V, W)$ denotes the set of all linear transformations $T : V \to W$, and $\mathrm{Mat}_{m\times n}(k)$ denotes the set of all $m \times n$ matrices with entries in k, then the function $T \mapsto {}_Y[T]_X$ is a bijection $\mathrm{Hom}_k(V, W) \to \mathrm{Mat}_{m\times n}(k)$.*

Proof. Given a matrix A, its columns define vectors in W; in more detail, if the jth column of A is $(a_{1j}, \ldots, a_{mj})$, define $z_j = \sum_{i=1}^{m} a_{ij} w_i$. By Theorem 3.92, there exists a linear transformation $T \colon V \to W$ with $T(v_j) = z_j$ and ${}_Y[T]_X = A$. Therefore, μ is surjective.

To see that μ is injective, suppose that ${}_Y[T]_X = A = {}_Y[S]_X$. Since the columns of A determine $T(v_j)$ and $S(v_j)$ for all j, Corollary 3.93 gives $S = T$. •

The next proposition shows where the definition of matrix multiplication comes from: the product of two matrices is the matrix of a composite.

Proposition 3.98. *Let $T \colon V \to W$ and $S \colon W \to U$ be linear transformations. Choose bases $X = x_1, \ldots, x_n$ of V, $Y = y_1, \ldots, y_m$ of W, and $Z = z_1, \ldots, z_\ell$ of U. Then*

$$ {}_Z[S \circ T]_X = \big({}_Z[S]_Y\big)\big({}_Y[T]_X\big). $$

Proof. Let ${}_Y[T]_X = [a_{ij}]$, so that $T(x_j) = \sum_p a_{pj} y_p$, and let ${}_Z[S]_Y = [b_{qp}]$, so that $S(y_p) = \sum_q b_{qp} z_q$. Then

$$ \begin{aligned} ST(x_j) = S(T(x_j)) &= S\Big(\sum_p a_{pj} y_p\Big) \\ &= \sum_p a_{pj} S(y_p) = \sum_p \sum_q a_{pj} b_{qp} z_q = \sum_q c_{qj} z_q, \end{aligned} $$

where $c_{qj} = \sum_p b_{qp} a_{pj}$. Therefore,

$$ {}_Z[ST]_X = [c_{qj}] = {}_Z[S]_Y \, {}_Y[T]_X. \quad \bullet $$

Corollary 3.99. *Matrix multiplication is associative.*

Proof. Let A be an $m \times n$ matrix, let B be an $n \times p$ matrix, and let C be a $p \times q$ matrix. By Theorem 3.92, there are linear transformations

$$ k^q \xrightarrow{T} k^p \xrightarrow{S} k^n \xrightarrow{R} k^m $$

with $C = [T]$, $B = [S]$, and $A = [R]$.

Then

$$ [R \circ (S \circ T)] = [R][S \circ T] = [R]([S][T]) = A(BC). $$

On the other hand,

$$ [(R \circ S) \circ T] = [R \circ S][T] = ([R][S])[T] = (AB)C. $$

Since composition of functions is associative,

$$ R \circ (S \circ T) = (R \circ S) \circ T, $$

and so

$$ A(BC) = [R \circ (S \circ T)] = [(R \circ S) \circ T] = (AB)C. \quad \bullet $$

We can prove Corollary 3.99 directly, although it is rather tedious, but the connection with composition of linear transformations is the real reason why matrix multiplication is associative.

Corollary 3.100. *Let $T : V \to W$ be a linear transformation of vector spaces V over a field k, and let X and Y be bases of V and W, respectively. If T is nonsingular, then the matrix of T^{-1} is the inverse of the matrix of T:*

$$_X[T^{-1}]_Y = ({}_Y[T]_X)^{-1}.$$

Proof. $I = {}_Y[1_W]_Y = {}_Y[T]_X\,{}_X[T^{-1}]_Y$ and $I = {}_X[1_V]_X = {}_X[T^{-1}]_Y\,{}_Y[T]_X$. •

The next corollary determines all the matrices arising from the same linear transformation.

Corollary 3.101. *Let $T : V \to V$ be a linear transformation on a vector space V over a field k. If X and Y are bases of V, then there is a nonsingular matrix P with entries in k so that*

$$_Y[T]_Y = P\big({}_X[T]_X\big)P^{-1}.$$

Conversely, if $B = PAP^{-1}$, where B, A, and P are $n \times n$ matrices with entries in k and P is nonsingular, then there is a linear transformation $T : k^n \to k^n$ and bases X and Y of k^n such that $B = {}_Y[T]_Y$ and $A = {}_X[T]_X$.

Proof. The first statement follows from Proposition 3.98 and associativity:

$$_Y[T]_Y = {}_Y[1_V T 1_V]_Y = ({}_Y[1_V]_X)({}_X[T]_X)({}_X[1_V]_Y).$$

Set $P = {}_Y[1_V]_X$, and note that Corollary 3.100 gives $P^{-1} = {}_X[1_V]_Y$.

For the converse, let $E = e_1, \ldots, e_n$ be the standard basis of k^n, and define $T : k^n \to k^n$ by $T(e_j) = Ae_j$ (remember that vectors in k^n are column vectors, so that Ae_j is matrix multiplication; indeed, Ae_j is the jth column of A). It follows that $A = {}_E[T]_E$. Now define a basis $Y = y_1, \ldots, y_n$ by $y_j = P^{-1}e_j$; that is, the vectors in Y are the columns of P^{-1}. Note that Y is a basis because P^{-1} is nonsingular. It suffices to prove that $B = {}_Y[T]_Y$; that is, $T(y_j) = \sum_i b_{ij} y_i$, where $B = [b_{ij}]$.

$$\begin{aligned}
T(y_j) &= Ay_j \\
&= AP^{-1}e_j \\
&= P^{-1}Be_j \\
&= P^{-1}\sum_i b_{ij}e_i \\
&= \sum_i b_{ij}P^{-1}e_i \\
&= \sum_i b_{ij}y_i \quad \bullet
\end{aligned}$$

Definition. Two $n \times n$ matrices B and A with entries in a field k are ***similar*** if there is a nonsingular matrix P with entries in k with $B = PAP^{-1}$.

Corollary 3.101 says that two matrices arise from the same linear transformation on a vector space V (from different choices of basis) if and only if they are similar. In Chapter 9, we will see how to determine whether two given matrices are similar.

Just as for group homomorphisms and ring homomorphisms, we can define the kernel and image of linear transformations.

Definition. If $T: V \to W$ is a linear transformation, then the ***kernel*** (or the ***null space***) of T is

$$\ker T = \{v \in V : T(v) = 0\},$$

and the ***image*** of T is

$$\operatorname{im} T = \{w \in W : w = T(v) \text{ for some } v \in V\}.$$

As in Proposition 3.94, an $m \times n$ matrix A with entries in a field k determines a linear transformation $k^n \to k^m$, namely, $y \mapsto Ay$, where y is an $n \times 1$ column vector. The kernel of this linear transformation is usually called the *solution space* of A [see Example 3.70(iv)].

The proof of the next proposition is routine.

Proposition 3.102. *Let $T: V \to W$ be a linear transformation.*

(i) $\ker T$ *is a subspace of V and* $\operatorname{im} T$ *is a subspace of W.*

(ii) T *is injective if and only if* $\ker T = \{0\}$.

We can now interpret the fact that a homogeneous system over a field k with r equations in n unknowns has a nontrivial solution if $r < n$. If A is the $r \times n$ coefficient matrix of the system, then $\varphi: x \mapsto Ax$ is a linear transformation $\varphi: k^n \to k^r$. If there is only the trivial solution, then $\ker \varphi = \{0\}$, so that k^n is isomorphic to a subspace of k^r, contradicting Corollary 3.90(i).

Lemma 3.103. *Let $T: V \to W$ be a linear transformation.*

(i) *If T is nonsingular, then for every basis $X = v_1, v_2, \ldots, v_n$ of V, we have $T(X) = T(v_1), T(v_2), \ldots, T(v_n)$ a basis of W.*

(ii) *Conversely, if there exists some basis $X = v_1, v_2, \ldots, v_n$ of V for which $T(X) = T(v_1), T(v_2), \ldots, T(v_n)$ is a basis of W, then T is nonsingular.*

Proof. (i) If $\sum c_i T(v_i) = 0$, then $T(\sum c_i v_i) = 0$, and so $\sum c_i v_i \in \ker T = \{0\}$. Hence each $c_i = 0$, because X is linearly independent. If $w \in W$, then the surjectivity of T provides $v \in V$ with $w = T(v)$. But $v = \sum a_i v_i$, and so $w = T(v) = T(\sum a_i v_i) = \sum a_i T(v_i)$. Therefore, $T(X)$ is a basis of W.

(ii) Let $w \in W$. Since $T(v_1), \ldots, T(v_n)$ is a basis of W, we have $w = \sum c_i T(v_i) = T(\sum c_i v_i)$, and so T is surjective. If $\sum c_i v_i \in \ker T$, then $\sum c_i T(v_i) = 0$, and so linear independence gives all $c_i = 0$; hence, $\sum c_i v_i = 0$ and $\ker T = \{0\}$. Therefore, T is nonsingular. •

Theorem 3.104. *If V is an n-dimensional vector space over a field k, then V is isomorphic to k^n.*

Proof. Choose a basis $v_1, \ldots, v_n$ of V. If $e_1, \ldots, e_n$ is the standard basis of k^n, then Theorem 3.92 says that there is a linear transformation $T : V \to k^n$ with $T(v_i) = e_i$ for all i; by Lemma 3.103, T is nonsingular. •

Theorem 3.104 does more than say that every finite-dimensional vector space is essentially the familiar vector space of all n-tuples. It says that a choice of basis in V is tantamount to a choice of coordinate set for each vector in V. We want the freedom to change coordinates because the usual coordinates may not be the most convenient ones for a given problem, as the reader has probably seen (in a calculus course) when rotating axes to simplify the equation of a conic section.

Corollary 3.105. *Two finite-dimensional vector spaces V and W over a field k are isomorphic if and only if* $\dim(V) = \dim(W)$.

Remark. In Theorem 6.51, we will see that this corollary remains true for infinite-dimensional vector spaces. ◂

Proof. Assume that there is a nonsingular $T : V \to W$. If $X = v_1, \ldots, v_n$ is a basis of V, then Lemma 3.103 says that $T(v_1), \ldots, T(v_n)$ is a basis of W. Therefore, $\dim(W) = |X| = \dim(V)$.

If $n = \dim(V) = \dim(W)$, then there are isomorphisms $T : V \to k^n$ and $S : W \to k^n$, by Theorem 3.104. It follows that the composite $S^{-1}T : V \to W$ is nonsingular. •

Proposition 3.106. *Let V be a finite-dimensional vector space with* $\dim(V) = n$*, and let $T : V \to V$ be a linear transformation. The following statements are equivalent:*

(i) *T is an isomorphism;*

(ii) *T is surjective;*

(iii) *T is injective.*

Proof. (i) ⇒ (ii) This implication is obvious.

(ii) ⇒ (iii) Let $v_1, \ldots, v_n$ be a basis of V. Since T is surjective, there are vectors $u_1, \ldots, u_n$ with $Tu_i = v_i$ for all i. We claim that $u_1, \ldots, u_n$ is linearly independent. If there are scalars $c_1, \ldots, c_n$, not all zero, with $\sum c_i u_i = 0$, then we obtain a dependency relation

$0 = \sum c_i T(u_i) = \sum c_i v_i$, a contradiction. By Corollary 3.89(ii), $u_1, \ldots, u_n$ is a basis of V. To show that T is injective, it suffices to show that $\ker T = \{0\}$. Suppose that $T(v) = 0$. Now $v = \sum c_i u_i$, and so $0 = T \sum c_i u_i = \sum c_i v_i$; hence, linear independence of $v_1, \ldots, v_n$ gives all $c_i = 0$, and so $v = 0$. Therefore, T is injective.

(iii) ⇒ (i) Let $v_1, \ldots, v_n$ be a basis of V. If $c_1, \ldots, c_n$ are scalars, not all 0, then $\sum c_i v_i \neq 0$, for a basis is linearly independent. Since T is injective, it follows that $\sum c_i T v_i \neq 0$, and so $T v_1, \ldots, T v_n$ is linearly independent. Therefore, Lemma 3.103(ii) shows that T is an isomorphism. •

Recall that an $n \times n$ matrix A with entries in a field k is *nonsingular* if there is a matrix B with entries in k (its *inverse*), with $AB = I = BA$. The next corollary shows that "one-sided inverses" are enough.

Corollary 3.107. *If A and B are $n \times n$ matrices with $AB = I$, then $BA = I$. Therefore, A is nonsingular with inverse B.*

Proof. There are linear transformations $T, S \colon k^n \to k^n$ with $[T] = A$ and $[S] = B$, and $AB = I$ gives

$$[TS] = [T][S] = [1_{k^n}].$$

Since $T \mapsto [T]$ is a bijection, by Proposition 3.97, it follows that $TS = 1_{k^n}$. By Proposition 1.47, T is a surjection and S is an injection. But Proposition 3.106 says that both T and S are isomorphisms, so that $S = T^{-1}$ and $TS = 1_{k^n} = ST$. Therefore, $I = [ST] = [S][T] = BA$, as desired. •

Definition. The set of all nonsingular $n \times n$ matrices with entries in k is denoted by $\mathrm{GL}(n, k)$.

Now that we have proven associativity, it is easy to prove that $\mathrm{GL}(n, k)$ is a group under matrix multiplication.

A choice of basis gives an isomorphism between the general linear group and the group of nonsingular matrices.

Proposition 3.108. *Let V be an n-dimensional vector space over a field k, and let $X = v_1, \ldots, v_n$ be a basis of V. Then $\mu \colon \mathrm{GL}(V) \to \mathrm{GL}(n, k)$, defined by $T \mapsto [T] = {}_X[T]_X$, is an isomorphism.*

Proof. By Proposition 3.97, the function $\mu' \colon T \mapsto [T] = {}_X[T]_X$ is a bijection

$$\mathrm{Hom}_k(V, V) \to \mathrm{Mat}_n(k),$$

where $\mathrm{Hom}_k(V, V)$ denotes the set of all linear transformations on V and $\mathrm{Mat}_n(k)$ denotes the set of all $n \times n$ matrices with entries in k. Moreover, Proposition 3.98 says that $[TS] = [T][S]$ for all $T, S \in \mathrm{Hom}_k(V, V)$.

If $T \in \mathrm{GL}(V)$, then $[T]$ is a nonsingular matrix, by Corollary 3.100; that is, if μ is the restriction of μ', then $\mu \colon \mathrm{GL}(V) \to \mathrm{GL}(n, k)$ is an injective homomorphism.

It remains to prove that μ is surjective. If $A \in \mathrm{GL}(n, k)$, then $A = [T]$ for some $T \colon V \to V$. It suffices to show that T is an isomorphism; that is, $T \in \mathrm{GL}(V)$. Since $[T]$ is a nonsingular matrix, there is a matrix B with $[T]B = I$. Now $B = [S]$ for some $S \colon V \to V$, and

$$[TS] = [T][S] = I = [1_V].$$

Therefore, $TS = 1_V$, since μ is a bijection, and so $T \in \mathrm{GL}(V)$, by Corollary 3.107. •

The center of the general linear group is easily identified; we now generalize Exercise 2.56 on page 81.

Definition. A linear transformation $T \colon V \to V$ is a ***scalar transformation*** if there is $c \in k$ with $T(v) = cv$ for all $v \in V$; that is, $T = c1_V$. A ***scalar matrix*** is a matrix of the form cI, where $c \in k$ and I is the identity matrix.

A scalar transformation $T = c1_V$ is nonsingular if and only if $c \neq 0$ (its inverse is $c^{-1}1_V$).

Corollary 3.109.

(i) *The center of the group* $\mathrm{GL}(V)$ *consists of all the nonsingular scalar transformations.*

(ii) *The center of the group* $\mathrm{GL}(n, k)$ *consists of all the nonsingular scalar matrices.*

Proof. (i) If $T \in \mathrm{GL}(V)$ is not scalar, then Example 3.96(ii) shows that there exists $v \in V$ with $v, T(v)$ linearly independent. By Proposition 3.87, there is a basis $v, T(v), u_3, \ldots, u_n$ of V. It is easy to see that $v, v + T(v), u_3, \ldots, u_n$ is also a basis of V, and so there is a nonsingular linear transformation S with $S(v) = v$, $S(T(v)) = v + T(v)$, and $S(u_i) = u_i$ for all i. Now S and T do not commute, for $ST(v) = v + T(v)$ while $TS(v) = T(v)$. Therefore, T is not in the center of $\mathrm{GL}(V)$.

(ii) If $f : G \to H$ is any group isomorphism between groups G and H, then $f(Z(G)) = Z(H)$. In particular, if $T = c1_V$ is a nonsingular scalar transformation, then $[T]$ is in the center of $\mathrm{GL}(n, k)$. But it is easily checked that $[T] = cI$ is a scalar matrix. •

EXERCISES

3.75 Let V and W be vector spaces over a field k, and let $S, T : V \to W$ be linear transformations.

(i) If V and W are finite-dimensional, prove that

$$\dim(\mathrm{Hom}_k(V, W)) = \dim(V)\dim(W).$$

(ii) The ***dual space*** V^* of a vector space V over k is defined by

$$V^* = \mathrm{Hom}_k(V, k).$$

If $\dim(V) = n$, prove that $\dim(V^*) = n$, and hence that $V^* \cong V$.

(iii) If $X = v_1, \ldots, v_n$ is a basis of V, define $\delta_1, \ldots, \delta_n \in V^*$ by

$$\delta_i(v_j) = \begin{cases} 0 & \text{if } j \neq i \\ 1 & \text{if } j = i. \end{cases}$$

Prove that $\delta_1, \ldots, \delta_n$ is a basis of V^* (it is called the ***dual basis*** arising from $v_1, \ldots, v_n$).

3.76 If $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$, define $\det(A) = ad - bc$. If V is a vector space with basis $X = v_1, v_2$, define $T : V \to V$ by $T(v_1) = av_1 + bv_2$ and $T(v_2) = cv_1 + dv_2$. Prove that T is a nonsingular linear transformation if and only if $\det({}_X[T]_X) \neq 0$.
Hint. You may assume the following (easily proved) fact of linear algebra: Given a system of linear equations with coefficients in a field,

$$ax + by = p$$
$$cx + dy = q,$$

then there exists a unique solution if and only if $ad - bc \neq 0$.

3.77 Let U be a subspace of a vector space V.

(i) Prove that the ***natural map*** $\pi : V \to V/U$, given by $v \mapsto v + U$, is a linear transformation with kernel U. (Quotient spaces were defined in Exercise 3.71 on page 170.)

(ii) State and prove the ***first isomorphism theorem*** for vector spaces.
Hint. Here is the statement. If $f : V \to W$ is a linear transformation with $\ker f = U$, then U is a subspace of V and there is an isomorphism $\varphi : V/U \cong \operatorname{im} f$, namely, $\varphi(v + U) = f(v)$.

3.78 Let V be a finite-dimensional vector space over a field k, and let $\mathcal{B}$ denote the family of all the bases of V. Prove that $\mathcal{B}$ is a transitive $\mathrm{GL}(V)$-set.
Hint. Use Theorem 3.92.

3.79 (i) If U and W are subspaces of a vector space V such that $U \cap W = \{0\}$ and $U + W = V$, prove that $V \cong U \oplus W$ (see the definition of direct sum on page 171).

(ii) A subspace U of a vector space V is a ***direct summand*** if there is a subspace W of V with $U \cap W = \{0\}$ and $U + W = V$. If V is a finite-dimensional vector space over a field k, prove that every subspace U is a direct summand.
Hint. Take a basis X of U, extend it to a basis X' of V, and define $W = \langle X' - X \rangle$.

3.80 If $T : V \to W$ is a linear transformation between vector spaces over a field k, define

$$\textbf{\textit{rank}}(T) = \dim(\operatorname{im} T).$$

(i) Regard the columns of an $m \times n$ matrix A as m-tuples, and define the ***column space*** of A to be the subspace of k^m spanned by the columns; define ***rank***(A) to be the dimension of the column space. If $T : k^n \to k^m$ is the linear transformation defined by $T(X) = AX$, where X is an $n \times 1$ vector, prove that

$$\operatorname{rank}(A) = \operatorname{rank}(T).$$

(ii) If A is an $m \times n$ matrix and B is an $p \times m$ matrix, prove that

$$\operatorname{rank}(BA) \leq \operatorname{rank}(A).$$

(iii) Prove that similar $n \times n$ matrices have the same rank.

3.8 QUOTIENT RINGS AND FINITE FIELDS

Let us return to commutative rings. The fundamental theorem of algebra (Theorem 4.49) states that every nonconstant polynomial in $\mathbb{C}[x]$ is a product of linear polynomials in $\mathbb{C}[x]$, that is, $\mathbb{C}$ contains all the roots of every polynomial in $\mathbb{C}[x]$. We are going to prove a "local" analog of the fundamental theorem of algebra for polynomials over an arbitrary field k: Given a polynomial $f(x) \in k[x]$, then there is some field K containing k that also contains all the roots of $f(x)$ (we call this a local analog for even though the larger field K contains all the roots of the polynomial $f(x)$, it may not contain roots of other polynomials in $k[x]$). The main idea behind the construction of K involves quotient rings, a construction akin to quotient groups.

Let I be an ideal in a commutative ring R. If we forget the multiplication, then I is a subgroup of the additive group R; since R is an abelian group, the subgroup I is necessarily normal, and so the quotient group R/I is defined, as is the natural map $\pi \colon R \to R/I$ given by $\pi(a) = a + I$. Recall Lemma 2.40(i), which we now write in additive notation: $a + I = b + I$ in R/I if and only if $a - b \in I$.

Theorem 3.110. *If I is an ideal in a commutative ring R, then the additive abelian group R/I can be made into a commutative ring in such a way that the natural map $\pi \colon R \to R/I$ is a surjective ring homomorphism.*

Sketch of Proof. Define multiplication on the additive abelian group R/I by

$$(a + I)(b + I) = ab + I.$$

To see that this is a well-defined function $R/I \times R/I \to R/I$, assume that $a + I = a' + I$ and $b + I = b' + I$, that is, $a - a' \in I$ and $b - b' \in I$. We must show that $(a' + I)(b' + I) = a'b' + I = ab + I$, that is, $ab - a'b' \in I$. But

$$\begin{aligned} ab - a'b' &= ab - a'b + a'b - a'b' \\ &= (a - a')b + a'(b - b') \in I, \end{aligned}$$

as desired.

To verify that R/I is a commutative ring, it now suffices to show associativity and commutativity of multiplication, distributivity, and that one is $1 + I$. Proofs of these properties are routine, for they are inherited from the corresponding property in R. For example, multiplication in R/I is commutative because

$$(a + I)(b + I) = ab + I = ba + I = (b + I)(a + I).$$

Rewriting the equation $(a + I)(b + I) = ab + I$ using the definition of π, namely, $a + I = \pi(a)$, gives $\pi(a)\pi(b) = \pi(ab)$. Since $\pi(1) = 1 + I$, it follows that π is a ring homomorphism. Finally, π is surjective because $a + I = \pi(a)$. •

Definition. The commutative ring R/I constructed in Theorem 3.110 is called the ***quotient ring***[14] of R modulo I (briefly, R mod I).

[14] Presumably, *quotient rings* are so called in analogy with quotient groups.

We saw in Example 2.68 that the additive abelian group $\mathbb{Z}/(m)$ is identical to $\mathbb{I}_m$. They have the same elements: the coset $a+(m)$ and the congruence class $[a]$ are the same subset of $\mathbb{Z}$; they have the same addition:

$$a + (m) + b + (m) = a + b + (m) = [a + b] = [a] + [b].$$

We can now see that the quotient ring $\mathbb{Z}/(m)$ coincides with the commutative ring $\mathbb{I}_m$, for the two multiplications coincide as well:

$$(a + (m))(b + (m)) = ab + (m) = [ab] = [a][b].$$

We can now prove a converse to Proposition 3.50.

Corollary 3.111. *If I is an ideal in a commutative ring R, then there are a commutative ring A and a ring homomorphism $\pi : R \to A$ with $I = \ker \pi$.*

Proof. If we forget the multiplication, then the natural map $\pi : R \to R/I$ is a homomorphism between additive groups and, by Corollary 2.69,

$$I = \ker \pi = \{r \in R : \pi(a) = 0 + I = I\}.$$

Now remember the multiplication: $(a + I)(b + I) = ab + I$; that is, $\pi(a)\pi(b) = \pi(ab)$. Therefore, π is a ring homomorphism, and $\ker \pi$ is equal to I whether the function π is regarded as a ring homomorphism or as a homomorphism of additive groups. •

Theorem 3.112 (First Isomorphism Theorem). *If $f : R \to A$ is a homomorphism of rings, then* $\ker f$ *is an ideal in R,* $\operatorname{im} f$ *is a subring of A, and*

$$R/\ker f \cong \operatorname{im} f.$$

Proof. Let $I = \ker f$. We have already seen, in Proposition 3.50, that I is an ideal in R and that $\operatorname{im} f$ is a subring of A.

If we forget the multiplication in the rings, then the proof of Theorem 2.70 shows that the function $\varphi \colon R/I \to A$, given by $\varphi(r + I) = f(r)$, is an isomorphism of additive groups. Since $\varphi(1+I) = f(1) = 1$, it now suffices to prove that φ preserves multiplication. But $\varphi\big((r + I)(s + I)\big) = \varphi(rs + I) = f(rs) = f(r)f(s) = \varphi(r + I)\varphi(s + I)$. Therefore, φ is a ring isomorphism. •

For rings as for groups, the first isomorphism theorem creates an isomorphism from a homomorphism once we know its kernel and image. It also says that there is no significant difference between a quotient ring and the image of a homomorphism. There are analogs for commutative rings of the second and third isomorphism theorems for groups (see Exercise 3.82 on page 196 for the third isomorphism theorem; the second isomorphism theorem is better stated in the context of modules; see Theorem 7.9), but they are less useful for rings than are their group analogs. However, there is a useful analog of the correspondence theorem, which we will prove later (see Proposition 6.1).

Definition. If k is a field, the intersection of all the subfields of k is called the ***prime field*** of k.

Every subfield of $\mathbb{C}$ contains $\mathbb{Q}$, and so the prime field of $\mathbb{C}$ and of $\mathbb{R}$ is $\mathbb{Q}$. The prime field of a finite field is just the integers mod p, as we show next.

Notation. From now on, we will denote $\mathbb{I}_p$ by $\mathbb{F}_p$ when we are regarding it as a field.

Borrowing terminology from group theory, call the intersection of all the subfields of a field containing a subset X the *subfield generated by* X; it is the smallest subfield containing X in the sense that if F is any subfield containing X, then F contains the subfield generated by X. The prime field is the subfield generated by 1, and the prime field of $\mathbb{F}_p(x)$ is $\mathbb{F}_p$.

Proposition 3.113. *If k is a field, then its prime field is isomorphic to $\mathbb{Q}$ or to $\mathbb{F}_p$ for some prime p.*

Proof. Consider the ring homomorphism $\chi \colon \mathbb{Z} \to k$, defined by $\chi(n) = n\varepsilon$, where we denote the *one* in k by ε. Since every ideal in $\mathbb{Z}$ is principal, there is an integer m with $\ker \chi = (m)$. If $m = 0$, then χ is an injection, and so there is an isomorphic copy of $\mathbb{Z}$ that is a subring of k. By Exercise 3.47(ii) on page 150, there is a field $Q \cong \text{Frac}(\mathbb{Z}) = \mathbb{Q}$ with $\text{im}\, \chi \subseteq Q \subseteq k$. Now Q is the prime field of k, for every subfield of k contains 1, hence contains $\text{im}\, \chi$, and hence it contains Q, for $Q \cong \mathbb{Q}$ has no proper subfields. If $m \neq 0$, the first isomorphism theorem gives $\mathbb{I}_m = \mathbb{Z}/(m) \cong \text{im}\, \chi \subseteq k$. Since k is a field, $\text{im}\, \chi$ is a domain, and so Proposition 3.6 gives m prime. If we now write p instead of m, then $\text{im}\, \chi = \{0, \varepsilon, 2\varepsilon, \ldots, (p-1)\varepsilon\}$ is a subfield of k isomorphic to $\mathbb{F}_p$. Clearly, $\text{im}\, \chi$ is the prime field of k, for every subfield contains ε, hence contains $\text{im}\, \chi$. •

This last result is the first step in classifying different types of fields.

Definition. A field k has ***characteristic* 0** if its prime field is isomorphic to $\mathbb{Q}$; a field k has ***characteristic p*** if its prime field is isomorphic to $\mathbb{F}_p$ for some prime p.

The fields $\mathbb{Q}$, $\mathbb{R}$, $\mathbb{C}$ have characteristic 0, as does any subfield of them; every finite field has characteristic p for some prime p, as does $\mathbb{F}_p(x)$, the ring of all rational functions over $\mathbb{F}_p$.

Proposition 3.114. *If k is a field of characteristic $p > 0$, then $pa = 0$ for all $a \in k$.*

Proof. Since k has characteristic p, we have $p \cdot 1 = 0$, where 1 is the *one* in k. The result now follows from Proposition 3.2(v). •

Proposition 3.115. *If k is a finite field, then $|k| = p^n$ for some prime p and some $n \geq 1$.*

Proof. The prime field P of k cannot be the infinite field $\mathbb{Q}$, and so $P \cong \mathbb{F}_p$ for some prime p. Now k is a vector space over P, and so it is a vector space over $\mathbb{F}_p$. Clearly, k is finite-dimensional, and if $\dim_{\mathbb{F}_p}(k) = n$, then $|k| = p^n$. •

Remark. Here is a proof of the last proposition using group theory. Assume that k is a finite field whose order $|k|$ is divisible by distinct primes p and q. By Proposition 2.78, Cauchy's theorem for abelian groups, there are elements a and b in k having orders p and q, respectively. If ε denotes *one* in k, then the elements $p\varepsilon$ (the sum of ε with itself p times) and $q\varepsilon$ satisfy $(p\varepsilon)a = 0$ and $(q\varepsilon)b = 0$. Since k is a field, it is a domain, and so

$$p\varepsilon = 0 = q\varepsilon.$$

But $(p, q) = 1$, so there are integers s and t with $sp+tq = 1$. Hence, $\varepsilon = s(p\varepsilon)+t(q\varepsilon) = 0$, and this is a contradiction. Therefore, $|k|$ has only one prime divisor, say, p, and so $|k|$ is a power of p. ◂

Proposition 3.116. *If k is a field and $I = (p(x))$, where $p(x)$ is a nonzero polynomial in $k[x]$, then the following are equivalent: $p(x)$ is irreducible; $k[x]/I$ is a field; $k[x]/I$ is a domain.*

Proof. Assume that $p(x)$ is irreducible. Note that $I = (p(x))$ is a proper ideal, so that the *one* in $k[x]/I$, namely, $1 + I$, is not zero. If $f(x) + I \in k[x]/I$ is nonzero, then $f(x) \notin I$, that is, $f(x)$ is not a multiple of $p(x)$ or, to say it another way, $p \nmid f$. By Lemma 3.36, p and f are relatively prime, and so there are polynomials s and t with $sf + tp = 1$. Thus, $sf - 1 \in I$, and so $1 + I = sf + I = (s + I)(f + I)$. Therefore, every nonzero element of $k[x]/I$ has an inverse, and so $k[x]/I$ is a field.

Of course, every field is a domain.

If $k[x]/I$ is a domain. If $p(x)$ is not an irreducible polynomial in $k[x]$, there is a factorization $p(x) = g(x)h(x)$ in $k[x]$ with $\deg(g) < \deg(p)$ and $\deg(h) < \deg(p)$. It follows that neither $g(x) + I$ nor $h(x) + I$ is zero in $k[x]/I$. After all, the zero in $k[x]/I$ is $0 + I = I$, and $g(x) + I = I$ if and only if $g(x) \in I = (p(x))$; but if this were so, then $p(x) \mid g(x)$, giving the contradiction $\deg(p) \leq \deg(g)$. The product

$$(g(x) + I)(h(x) + I) = p(x) + I = I$$

is zero in the quotient ring, and this contradicts $k[x]/I$ being a domain. Therefore, $p(x)$ must be an irreducible polynomial. •

The structure of R/I can be rather complicated, but for special choices of R and I, the commutative ring R/I can be easily described. For example, when $p(x)$ is an irreducible polynomial, the following proposition gives a complete description of the field $k[x]/(p(x))$.

Proposition 3.117. *Let k be a field, let $p(x) \in k[x]$ be a monic irreducible polynomial of degree d, let $K = k[x]/I$, where $I = (p(x))$, and let $\beta = x + I \in K$.*

(i) *K is a field and $k' = \{a + I : a \in k\}$ is a subfield of K isomorphic to k. Therefore, if k' is identified with k, then k is a subfield of K.*

(ii) *β is a root of $p(x)$ in K.*

(iii) *If $g(x) \in k[x]$ and β is a root of $g(x)$, then $p(x) \mid g(x)$ in $k[x]$.*

(iv) *$p(x)$ is the unique monic irreducible polynomial in $k[x]$ having β as a root.*

(v) *The list $1, \beta, \beta^2, \ldots, \beta^{d-1}$ is a basis of K as a vector space over k, and so $\dim_k(K) = d$.*

Proof. (i) The quotient ring $K = k[x]/I$ is a field, by Proposition 3.116, because $p(x)$ is irreducible. It is easy to see, using Corollary 3.53, that the restriction of the natural map, $\varphi = \pi|k \colon k \to K$, defined by $\varphi(a) = a + I$, is an isomorphism from $k \to k'$.

(ii) Let $p(x) = a_0 + a_1 x + \cdots + a_{d-1}x^{d-1} + x^d$, where $a_i \in k$ for all i. In $K = k[x]/I$, we have

$$\begin{aligned} p(\beta) &= (a_0 + I) + (a_1 + I)\beta + \cdots + (1 + I)\beta^d \\ &= (a_0 + I) + (a_1 + I)(x + I) + \cdots + (1 + I)(x + I)^d \\ &= (a_0 + I) + (a_1 x + I) + \cdots + (1x^d + I) \\ &= a_0 + a_1 x + \cdots + x^d + I \\ &= p(x) + I = I, \end{aligned}$$

because $p(x) \in I = (p(x))$. But $I = 0 + I$ is the zero element of $K = k[x]/I$, and so β is a root of $p(x)$.

(iii) If $p(x) \nmid g(x)$ in $k[x]$, then their gcd is 1, because $p(x)$ is irreducible. Therefore, there are $s(x), t(x) \in k[x]$ with $1 = s(x)p(x) + t(x)g(x)$. Since $k[x] \subseteq K[x]$, we may regard this as an equation in $K[x]$. Evaluating at β gives the contradiction $1 = 0$.

(iv) Let $h(x) \in k[x]$ be a monic irreducible polynomial having β as a root. By part (iii), we have $p(x) \mid h(x)$. Since $h(x)$ is irreducible, we have $h(x) = cp(x)$ for some constant c; since $h(x)$ and $p(x)$ are monic, we have $c = 1$ and $h(x) = p(x)$.

(v) Every element of K has the form $f(x) + I$, where $f(x) \in k[x]$. By the division algorithm, there are polynomials $q(x), r(x) \in k[x]$ with $f(x) = q(x)p(x) + r(x)$ and either $r(x) = 0$ or $\deg(r) < d = \deg(p)$. Since $f - r = qp \in I$, it follows that $f(x) + I = r(x) + I$. If $r(x) = b_0 + b_1 x + \cdots + b_{d-1}x^{d-1}$, where $b_i \in k$ for all i, then we see, as in the proof of part (ii), that $r(x) + I = b_0 + b_1\beta + \cdots + b_{d-1}\beta^{d-1}$. Therefore, $1, \beta, \beta^2, \ldots, \beta^{d-1}$ spans K.

To prove uniqueness, suppose that

$$b_0 + b_1\beta + \cdots + b_{d-1}\beta^{n-1} = c_0 + c_1\beta + \cdots + c_{d-1}\beta^{d-1}.$$

Define $g(x) \in k[x]$ by $g(x) = \sum_{i=0}^{d-1}(b_i - c_i)x^i$; if $g(x) = 0$, we are done. If $g(x) \neq 0$, then $\deg(g)$ is defined, and $\deg(g) < d = \deg(p)$. On the other hand, β is a root of $g(x)$, and so part (iii) gives $p(x) \mid g(x)$; hence, $\deg(p) \leq \deg(g)$, and this is a contradiction. It follows that $1, \beta, \beta^2, \ldots, \beta^{d-1}$ is a basis of K as a vector space over k, and this gives $\dim_k(K) = d$. •

Definition. If K is a field containing k as a subfield, then K is called a (field) ***extension*** of k, and we write "K/k is a field extension."[15]

An extension field K of a field k is a ***finite extension*** of k if K is a finite-dimensional vector space over k. The dimension of K, denoted by $[K : k]$, is called the ***degree*** of K/k.

Proposition 3.117(v) shows why $[K : k]$ is called the degree of the extension K/k.

Example 3.118.
The polynomial $x^2+1 \in \mathbb{R}[x]$ is irreducible, and so $K = \mathbb{R}[x]/(x^2+1)$ is a field extension $K/\mathbb{R}$ of degree 2. If β is a root of $x^2 + 1$, then $\beta^2 = -1$; moreover, every element of K has a unique expression of the form $a + b\beta$, where $a, b \in \mathbb{R}$. Clearly, this is another construction of $\mathbb{C}$ (which we have been viewing as the points in the plane equipped with a certain addition and multiplication).

Here is a natural way to construct an isomorphism $K \to \mathbb{C}$. Consider the evaluation map $\varphi\colon \mathbb{R}[x] \to \mathbb{C}$ given by $\varphi\colon f(x) \mapsto f(i)$. First, φ is surjective, for $a + ib = \varphi(a + bx) \in \operatorname{im}\varphi$. Second, $\ker\varphi = \{f(x) \in \mathbb{R}[x] : f(i) = 0\}$, the set of all polynomials in $\mathbb{R}[x]$ having i as a root. We know that $x^2 + 1 \in \ker\varphi$, so that $(x^2 + 1) \subseteq \ker\varphi$. For the reverse inclusion, take $g(x) \in \ker\varphi$. Now i is a root of $g(x)$, and so gcd $(g, x^2 + 1) \neq 1$ in $\mathbb{C}[x]$; therefore, gcd $(g, x^2 + 1) \neq 1$ in $\mathbb{R}[x]$. Irreducibility of $x^2 + 1$ in $\mathbb{R}[x]$ gives $x^2 + 1 \mid g(x)$, and so $g(x) \in (x^2 + 1)$, Therefore, $\ker\varphi = (x^2 + 1)$. The first isomorphism theorem now gives $\mathbb{R}[x]/(x^2 + 1) \cong \mathbb{C}$. ◂

The easiest way to multiply in $\mathbb{C}$ is to first treat i as a variable and then to impose the condition $i^2 = -1$. To compute $(a + bi)(c + di)$, first write $ac + (ad + bc)i + bdi^2$, and then observe that $i^2 = -1$. More generally, if β is a root of an irreducible $p(x) \in k[x]$, then the proper way to multiply

$$(b_0 + b_1\beta + \cdots + b_{n-1}\beta^{n-1})(c_0 + c_1\beta + \cdots + c_{n-1}\beta^{n-1})$$

in the quotient ring $k[x]/(p(x))$ is to regard the factors as polynomials in β, multiply them, and then impose the condition that $p(\beta) = 0$.

A first step in classifying fields involves their characteristic; that is, describing prime fields. A next step considers whether the elements are *algebraic* over the prime field.

Definition. Let K/k be a field extension. An element $\alpha \in K$ is ***algebraic*** over k if there is some nonzero polynomial $f(x) \in k[x]$ having α as a root; otherwise, α is ***transcendental*** over k. An extension K/k is ***algebraic*** if every $\alpha \in K$ is algebraic over k.

When a real number is called transcendental, it usually means that it is transcendental over $\mathbb{Q}$.

Proposition 3.119. *If K/k is a finite field extension, then K/k is an algebraic extension.*

[15]This notation should not be confused with the notation for a quotient ring, for a field K has no interesting ideals; in particular, if $k \subsetneq K$, then k is not an ideal in K.

Proof. By definition, K/k finite means that $[K : k] = n < \infty$; that is, K has dimension n as a vector space over k. By Corollary 3.88, the list of $n + 1$ vectors $1, \alpha, \alpha^2, \ldots, \alpha^n$ is dependent. Thus, there are $c_0, c_1, \ldots, c_n \in k$, not all 0, with $\sum c_i \alpha^i = 0$. Thus, the polynomial $f(x) = \sum c_i x^i$ is not the zero polynomial, and α is a root of $f(x)$. Therefore, α is algebraic over k. •

The converse of this last proposition is not true. We shall see, in Example 6.55, that the set $\mathbb{A}$ of all complex numbers algebraic over $\mathbb{Q}$ is an algebraic extension of $\mathbb{Q}$ that is not a finite extension.

Definition. If K/k is an extension and $\alpha \in K$, then $k(\alpha)$ is the intersection of all those subfields of K that contain k and α; we call $k(\alpha)$ the subfield of K obtained by ***adjoining*** α to k.

More generally, if A is a (possibly infinite) subset of K, define $k(A)$ to be the intersection of all the subfields of K that contain $k \cup A$; we call $k(A)$ the subfield of K obtained by ***adjoining*** A to k. In particular, if $A = \{z_1, \ldots, z_n\}$ is a finite subset, then we may denote $k(A)$ by $k(z_1, \ldots, z_n)$.

It is clear that $k(A)$ is the smallest subfield of K containing k and A; that is, if B is any subfield of K containing k and A, then $k(A) \subseteq B$.

We now show that the field $k[x]/(p(x))$, where $p(x) \in k[x]$ is irreducible, is intimately related to adjunction.

Theorem 3.120.

(i) *If K/k is an extension and $\alpha \in K$ is algebraic over k, then there is a unique monic irreducible polynomial $p(x) \in k[x]$ having α as a root. Moreover, if $I = (p(x))$, then $k[x]/I \cong k(\alpha)$; indeed, there exists an isomorphism*

$$\varphi : k[x]/I \to k(\alpha)$$

with $\varphi(x + I) = \alpha$ and $\varphi(c + I) = c$ for all $c \in k$.

(ii) *If $\alpha' \in K$ is another root of $p(x)$, then there is an isomorphism*

$$\theta : k(\alpha) \to k(\alpha')$$

with $\theta(\alpha) = \alpha'$ and $\theta(c) = c$ for all $c \in k$.

Proof. (i) Consider evaluation, the ring homomorphism $\varphi \colon k[x] \to K$ defined by

$$\varphi : f(x) \mapsto f(\alpha).$$

Now $\operatorname{im}\varphi$ is the subring of K consisting of all polynomials in α; that is, all elements of the form $f(\alpha)$ with $f(x) \in k[x]$. Now $\ker\varphi$ is the ideal in $k[x]$ consisting of all those $f(x) \in k[x]$ having α as a root. Since every ideal in $k[x]$ is a principal ideal, we have $\ker\varphi = (p(x))$ for some monic polynomial $p(x) \in k[x]$. But $k[x]/(p(x)) \cong \operatorname{im}\varphi$, which

is a domain, and so $p(x)$ is irreducible, by Proposition 3.116. This same proposition says that $k[x]/(p(x))$ is a field, and so the first isomorphism theorem gives $k[x]/(p(x)) \cong \operatorname{im}\varphi$; that is, $\operatorname{im}\varphi$ is a subfield of K containing k and α. Since every subfield of K that contains k and α must contain $\operatorname{im}\varphi$, we have $\operatorname{im}\varphi = k(\alpha)$. We have proved everything in the statement except the uniqueness of $p(x)$; but this now follows from Proposition 3.117(iv).

(ii) As in part (i), there are isomorphisms $\varphi\colon k[x]/I \to k(\alpha)$ and $\psi\colon k[x]/I \to k(\alpha')$ with $\varphi(c+I) = c$ and $\psi(c) = c+I$ for all $c \in k$; moreover, $\varphi\colon x+I \mapsto \alpha$ and $\psi\colon x+I \mapsto \alpha'$. The composite $\theta = \psi\varphi^{-1}$ is the desired isomorphism. •

Definition. If K/k is a field extension and $\alpha \in K$ is algebraic over k, then the unique monic irreducible polynomial $p(x) \in k[x]$ having α as a root is called the ***minimal polynomial*** of α over k, and it is denoted by

$$\operatorname{irr}(\alpha, k) = p(x).$$

The minimal polynomial $\operatorname{irr}(\alpha, k)$ does depend on k. For example, $\operatorname{irr}(i, \mathbb{R}) = x^2 + 1$, while $\operatorname{irr}(i, \mathbb{C}) = x - i$.

The following formula is quite useful, especially when proving a theorem by induction on degrees.

Theorem 3.121. *Let $k \subseteq E \subseteq K$ be fields, with E a finite extension of k and K a finite extension of E. Then K is a finite extension of k, and*

$$[K : k] = [K : E][E : k].$$

Proof. If $A = a_1, \ldots, a_n$ is a basis of E over k and if $B = b_1, \ldots, b_m$ is a basis of K over E, then it suffices to prove that a list X of all $a_i b_j$ is a basis of K over k.

To see that X spans K, take $u \in K$. Since B is a basis of K over E, there are scalars $\lambda_j \in E$ with $u = \sum_j \lambda_j b_j$. Since A is a basis of E over k, there are scalars $\mu_{ji} \in k$ with $\lambda_j = \sum_i \mu_{ji} a_i$. Therefore, $u = \sum_{ij} \mu_{ji} a_i b_j$, and X spans K over k.

To prove that X is linearly independent over k, assume that there are scalars $\mu_{ji} \in k$ with $\sum_{ij} \mu_{ji} a_i b_j = 0$. If we define $\lambda_j = \sum_i \mu_{ji} a_i$, then $\lambda_j \in E$ and $\sum_j \lambda_j b_j = 0$. Since B is linearly independent over E, it follows that

$$0 = \lambda_j = \sum_i \mu_{ji} a_i$$

for all j. Since A is linearly independent over k, it follows that $\mu_{ji} = 0$ for all j and i, as desired. •

There are several classical problems in euclidean geometry: trisecting an angle; *duplicating the cube* (given a cube with side length 1, construct a cube whose volume is 2); *squaring the circle* (given a circle of radius 1, construct a square whose area is equal to the area of the circle). In short, the problems ask whether geometric constructions can be made

using only a straightedge (ruler) and compass according to certain rules. Theorem 3.121 has a beautiful application in proving the unsolvability of these classical problems. For a discussion of these results, the reader may see my book, *A First Course in Abstract Algebra*, pages 332–344.

Example 3.122.
Let $f(x) = x^4 - 10x^2 + 1 \in \mathbb{Q}[x]$. If β is a root of $f(x)$, then the quadratic formula gives $\beta^2 = 5 \pm 2\sqrt{6}$. But the identity $a + 2\sqrt{ab} + b = \big(\sqrt{a} + \sqrt{b}\big)^2$ gives $\beta = \pm(\sqrt{2} + \sqrt{3})$. Similarly, $5 - 2\sqrt{6} = \big(\sqrt{2} - \sqrt{3}\big)^2$, so that the roots of $f(x)$ are

$$\sqrt{2} + \sqrt{3}, \quad -\sqrt{2} - \sqrt{3}, \quad \sqrt{2} - \sqrt{3}, \quad -\sqrt{2} + \sqrt{3}.$$

By Theorem 3.43, the only possible rational roots of $f(x)$ are ± 1, and so we have just proved that these roots are irrational.

We claim that $f(x)$ is irreducible in $\mathbb{Q}[x]$. If $g(x)$ is a quadratic factor of $f(x)$ in $\mathbb{Q}[x]$, then

$$g(x) = \big(x - a\sqrt{2} - b\sqrt{3}\big)\big(x - c\sqrt{2} - d\sqrt{3}\big),$$

where $a, b, c, d \in \{1, -1\}$. Multiplying,

$$g(x) = x^2 - \Big((a + c)\sqrt{2} + (b + d)\sqrt{3}\Big)x + 2ac + 3bd + (ad + bc)\sqrt{6}.$$

We check easily that $(a + c)\sqrt{2} + (b + d)\sqrt{3}$ is rational if and only if $a + c = 0 = b + d$; but these equations force $ad + bc \neq 0$, and so the constant term of $g(x)$ is not rational. Therefore, $g(x) \notin \mathbb{Q}[x]$, and so $f(x)$ is irreducible in $\mathbb{Q}[x]$. If $\beta = \sqrt{2} + \sqrt{3}$, then $f(x) = \operatorname{irr}(\beta, \mathbb{Q})$.

Consider the field $E = \mathbb{Q}(\beta) = \mathbb{Q}\big(\sqrt{2} + \sqrt{3}\big)$. There is a tower of fields $\mathbb{Q} \subseteq E \subseteq F$, where $F = \mathbb{Q}(\sqrt{2}, \sqrt{3})$, and so

$$[F : \mathbb{Q}] = [F : E][E : \mathbb{Q}],$$

by Theorem 3.121. Since $E = \mathbb{Q}(\beta)$ and β is a root of an irreducible polynomial of degree 4, namely, $f(x)$, we have $[E : \mathbb{Q}] = 4$. On the other hand,

$$[F : \mathbb{Q}] = [F : \mathbb{Q}\big(\sqrt{2}\big)][\mathbb{Q}\big(\sqrt{2}\big) : \mathbb{Q}].$$

Now $[\mathbb{Q}\big(\sqrt{2}\big) : \mathbb{Q}] = 2$, because $\sqrt{2}$ is a root of the irreducible quadratic $x^2 - 2$ in $\mathbb{Q}[x]$. We claim that $[F : \mathbb{Q}\big(\sqrt{2}\big)] \leq 2$. The field F arises by adjoining $\sqrt{3}$ to $\mathbb{Q}\big(\sqrt{2}\big)$; either $\sqrt{3} \in \mathbb{Q}\big(\sqrt{2}\big)$, in which case the degree is 1, or $x^2 - 3$ is irreducible in $\mathbb{Q}\big(\sqrt{2}\big)[x]$, in which case the degree is 2 (in fact, the degree is 2). It follows that $[F : \mathbb{Q}] \leq 4$, and so the equation $[F : \mathbb{Q}] = [F : E][E : \mathbb{Q}]$ gives $[F : E] = 1$; that is, $F = E$.

Let us note that F arises from $\mathbb{Q}$ by adjoining all the roots of $f(x)$, and it also arises from $\mathbb{Q}$ by adjoining all the roots of $g(x) = (x^2 - 2)(x^2 - 3)$. ◀

We now prove two important results: The first, due to L. Kronecker, says that if $f(x) \in k[x]$, where k is any field, then there is some larger field E that contains k and all the roots of $f(x)$; the second, due to E. Galois, constructs finite fields other than $\mathbb{F}_p$.

Theorem 3.123 (***Kronecker***). *If k is a field and $f(x) \in k[x]$, then there exists a field K containing k as a subfield and with $f(x)$ a product of linear polynomials in $K[x]$.*

Proof. The proof is by induction on $\deg(f)$. If $\deg(f) = 1$, then $f(x)$ is linear and we can choose $K = k$. If $\deg(f) > 1$, write $f(x) = p(x)g(x)$, where $p(x)$ is irreducible. Now Proposition 3.117(i) provides a field F containing k and a root z of $p(x)$. Hence, in $F[x]$, we have $p(x) = (x - z)h(x)$ and $f(x) = (x - z)h(x)g(x)$. By induction, there is a field K containing F (and hence k) so that $h(x)g(x)$, and hence $f(x)$, is a product of linear factors in $K[x]$. •

For the familiar fields $\mathbb{Q}$, $\mathbb{R}$, and $\mathbb{C}$, Kronecker's theorem offers nothing new. The *fundamental theorem of algebra*, first proved by Gauss in 1799 (completing earlier attempts of Euler and of Lagrange), says that every nonconstant $f(x) \in \mathbb{C}[x]$ has a root in $\mathbb{C}$; it follows, by induction on the degree of $f(x)$, that all the roots of $f(x)$ lie in $\mathbb{C}$; that is, $f(x) = a(x - r_1) \ldots (x - r_n)$, where $a \in \mathbb{C}$ and $r_j \in \mathbb{C}$ for all j. On the other hand, if $k = \mathbb{F}_p$ or $k = \mathbb{C}(x) = \mathrm{Frac}(\mathbb{C}[x])$, then the fundamental theorem does not apply; but Kronecker's theorem does apply to tell us, for any given $f(x)$, that there is always some larger field E that contains all the roots of $f(x)$. For example, there is some field containing $\mathbb{C}(x)$ and $\sqrt{x}$. There is a general version of the fundamental theorem that we give in Chapter 6: Every field k is a subfield of an ***algebraically closed*** field K, that is, K is a field containing k such that every $f(x) \in K[x]$ is a product of linear polynomials in $K[x]$. In contrast, Kronecker's theorem gives roots of just one polynomial at a time.

The definition of $k(A)$, the field obtained by adjoining a set A to k, assumes that A is a subset of a field extension K of k. In light of Kronecker's theorem, we may now speak of a field extension $k(z_1, \ldots, z_n)$ obtained by adjoining all the roots of some $f(x) \in k[x]$ without having to wonder whether such an extension K/k exists.

Definition. Let k be a subfield of a field K, and let $f(x) \in k[x]$. We say that $f(x)$ ***splits over K*** if

$$f(x) = a(x - z_1) \cdots (x - z_n),$$

where $z_1, \ldots, z_n$ are in K and $a \in k$ is nonzero.

If $f(x) \in k[x]$ is a polynomial, then a field extension E/k is called a ***splitting field*** of $f(x)$ ***over*** k if $f(x)$ splits over E, but $f(x)$ does not split over any proper subfield of E.

For example, consider $f(x) = x^2 + 1 \in \mathbb{Q}[x]$. The roots of $f(x)$ are $\pm i$, and so $f(x)$ splits over $\mathbb{C}$; that is, $f(x) = (x - i)(x + i)$ is a product of linear polynomials in $\mathbb{C}[x]$. However, $\mathbb{C}$ is not a splitting field over $\mathbb{Q}$, for $\mathbb{C}$ is not the *smallest* field containing $\mathbb{Q}$ and all the roots of $f(x)$. The splitting field of $f(x) \in k[x]$ depends on k as well as on $f(x)$: Here, the splitting field over $\mathbb{Q}$ is $\mathbb{Q}(i)$; the splitting field over $\mathbb{R}$ is $\mathbb{R}(i) = \mathbb{C}$.

In Example 3.122, we proved that $E = \mathbb{Q}(\sqrt{2} + \sqrt{3})$ is a splitting field of $f(x) = x^4 - 10x^2 + 1$, as well as a splitting field of $g(x) = (x^2 - 2)(x^2 - 3)$.

The existence of splitting fields is an easy consequence of Kronecker's theorem.

Corollary 3.124. *Let k be a field, and let $f(x) \in k[x]$. Then a splitting field of $f(x)$ over k exists.*

Proof. By Kronecker's theorem, there is a field extension K/k such that $f(x)$ splits in $K[x]$; say, $f(x) = a(x - \alpha_1) \cdots (x - \alpha_n)$. The subfield $E = k(\alpha_1, \ldots, \alpha_n)$ of K is a splitting field of $f(x)$ over k. •

Thus, a splitting field of $f(x) \in k[x]$ is the smallest subfield E of K containing k and all the roots of $f(x)$. The reason we say "a" splitting field instead of "the" splitting field is that the definition involves not only $f(x)$ and k, but the larger field K as well. Analysis of this technical point will enable us to prove Corollary 3.132: Any two finite fields with the same number of elements are isomorphic.

Example 3.125.
Let k be a field and let $E = k(y_1, \ldots, y_n)$ be the rational function field in n variables $y_1, \ldots, y_n$ over k; that is, $E = \mathrm{Frac}(k[y_1, \ldots, y_n])$, the fraction field of the ring of polynomials in n variables. The ***general polynomial of degree*** n over k is defined to be

$$f(x) = \prod_i (x - y_i) \in \mathrm{Frac}(k[y_1, \ldots, y_n])[x].$$

The coefficients of $f(x) = (x - y_1)(x - y_2) \cdots (x - y_n)$, which we denote by a_i, can be given explicitly [see Eqs. (1) on page 198] in terms of the y's. Notice that E is a splitting field of $f(x)$ over the field $K = k(a_0, \ldots, a_{n-1})$, for it arises from K by adjoining to it all the roots of $f(x)$, namely, all the y's. ◄

Here is another application of Kronecker's theorem.

Proposition 3.126. *Let p be a prime, and let k be a field. If $f(x) = x^p - c \in k[x]$ and α is a pth root of c (in some splitting field), then either $f(x)$ is irreducible in $k[x]$ or c has a pth root in k. In either case, if k contains the pth roots of unity, then $k(\alpha)$ is a splitting field of $f(x)$.*

Proof. By Kronecker's theorem, there exists a field extension K/k that contains all the roots of $f(x)$; that is, K contains all the pth roots of c. If $\alpha^p = c$, then every such root has the form $\omega\alpha$, where ω is a pth root of unity; that is, ω is a root of $x^p - 1$.

If $f(x)$ is not irreducible in $k[x]$, then there is a factorization $f(x) = g(x)h(x)$ in $k[x]$ with $g(x)$ a nonconstant polynomial with $d = \deg(g) < \deg(f) = p$. Now the constant term b of $g(x)$ is, up to sign, the product of some of the roots of $f(x)$:

$$\pm b = \alpha^d \omega,$$

where ω, which is a product of d pth roots of unity, is itself a pth root of unity. It follows that

$$(\pm b)^p = (\alpha^d \omega)^p = \alpha^{dp} = c^d.$$

But p being prime and $d < p$ forces $(d, p) = 1$; hence, there are integers s and t with $1 = sd + tp$. Therefore,

$$c = c^{sd+tp} = c^{sd} c^{tp} = (\pm b)^{ps} c^{tp} = [(\pm b)^s c^t]^p.$$

Therefore, c has a pth root in k.

If $\alpha \in K$ is a pth root of c, then $f(x) = \prod_{\omega}(x - \omega\alpha)$, where ω ranges over the pth roots of unity. Since we are now assuming that all ω lie in k, it follows that $k(\alpha)$ is a splitting field of $f(x)$. •

It follows, for every prime p, that $x^p - 2$ is irreducible in $\mathbb{Q}[x]$.

We are now going to construct the finite fields. My guess is that Galois knew that $\mathbb{C}$ can be constructed by adjoining a root of a polynomial, namely, $x^2 + 1$, to $\mathbb{R}$, and so it was natural for him to adjoin a root of a polynomial to $\mathbb{F}_p$. Note, however, that Kronecker's theorem was not proved until a half century after Galois's death.

Theorem 3.127 (*Galois*). *If p is a prime and n is a positive integer, then there is a field having exactly p^n elements.*

Proof. Write $q = p^n$, and consider the polynomial

$$g(x) = x^q - x \in \mathbb{F}_p[x].$$

By Kronecker's theorem, there is a field K containing $\mathbb{F}_p$ such that $g(x)$ is a product of linear factors in $K[x]$. Define

$$E = \{\alpha \in K : g(\alpha) = 0\};$$

thus, E is the set of all the roots of $g(x)$. Since the derivative $g'(x) = qx^{q-1} - 1 = p^n x^{q-1} - 1 = -1$ (see Exercise 3.23 on page 130), it follows that the $\gcd(g, g')$ is 1. By Exercise 3.37 on page 142, all the roots of $g(x)$ are distinct; that is, E has exactly $q = p^n$ elements.

We claim that E is a subfield of K, and this will complete the proof. If $a, b \in E$, then $a^q = a$ and $b^q = b$. Therefore, $(ab)^q = a^q b^q = ab$, and $ab \in E$. By Exercise 3.45 on page 149(iii), $(a - b)^q = a^q - b^q = a - b$, so that $a - b \in E$. Finally, if $a \neq 0$, then the cancellation law applied to $a^q = a$ gives $a^{q-1} = 1$, and so the inverse of a is a^{q-2} (which lies in E because E is closed under multiplication). •

We will soon see that any two finite fields with the same number of elements are isomorphic.

Recall Theorem 3.30: The multiplicative group of a finite field k is a cyclic group; a generator α of this group is called a ***primitive element***; that is, every nonzero element of k is a power of α.

Notation. Denote a finite field having $q = p^n$ elements (where p is a prime) by

$$\mathbb{F}_q.$$

Corollary 3.128. *For every prime p and every integer $n \geq 1$, there exists an irreducible polynomial $g(x) \in \mathbb{F}_p[x]$ of degree n. In fact, if α is a primitive element of $\mathbb{F}_{p^n}$, then its minimal polynomial $g(x) = \operatorname{irr}(\alpha, \mathbb{F}_p)$ has degree n.*

Remark. An easy modification of the proof replaces $\mathbb{F}_p$ by any finite field. ◂

Proof. Let $E/\mathbb{F}_p$ be a field extension with p^n elements, and let $\alpha \in E$ be a primitive element. Clearly, $\mathbb{F}_p(\alpha) = E$, for it contains every power of α, hence every nonzero element of E. By Theorem 3.120(i), $g(x) = \operatorname{irr}(\alpha, \mathbb{F}_p) \in \mathbb{F}_p[x]$ is an irreducible polynomial having α as a root. If $\deg(g) = d$, then Proposition 3.117(v) gives $[\mathbb{F}_p[x]/(g(x)) : \mathbb{F}_p] = d$; but $\mathbb{F}_p[x]/(g(x)) \cong \mathbb{F}_p(\alpha) = E$, by Theorem 3.120(i), so that $[E : \mathbb{F}_p] = n$. Therefore, $n = d$, and so $g(x)$ is an irreducible polynomial of degree n. •

This corollary can also be proved by counting. If $m = p_1^{e_1} \cdots p_n^{e_n}$, define the ***Möbius function*** by

$$\mu(m) = \begin{cases} 1 & \text{if } m = 1; \\ 0 & \text{if any } e_i > 1; \\ (-1)^n & \text{if } 1 = e_1 = e_2 = \cdots = e_n. \end{cases}$$

If N_n is the number of irreducible polynomials in $\mathbb{F}_p[x]$ of degree n, then

$$N_n = \frac{1}{n} \sum_{d|n} \mu(d) p^{n/d}.$$

An elementary proof can be found in G. J. Simmons, "The Number of Irreducible Polynomials of Degree n over GF(p)," *American Mathematical Monthly* 77 (1970), pages 743–745.

Example 3.129.
(i) In Exercise 3.14 on page 125, we constructed a field k with four elements:

$$k = \left\{ \begin{bmatrix} a & b \\ b & a+b \end{bmatrix} : a, b \in \mathbb{I}_2 \right\}.$$

On the other hand, we may construct a field of order 4 as the quotient $F = \mathbb{F}_2[x]/(q(x))$, where $q(x) \in \mathbb{F}_2[x]$ is the irreducible polynomial x^2+x+1. By Proposition 3.117(v), F is a field consisting of all $a + bz$, where $z = x + (q(x))$ is a root of $q(x)$ and $a, b \in \mathbb{I}_2$. Since $z^2+z+1 = 0$, we have $z^2 = -z-1 = z+1$; moreover, $z^3 = zz^2 = z(z+1) = z^2+z = 1$. It is now easy to see that there is a ring isomorphism $\varphi : k \to F$ with $\varphi\left(\begin{bmatrix} a & b \\ b & a+b \end{bmatrix}\right) = a + bz$.

(ii) According to the table in Example 3.35(ii) on page 137, there are three monic irreducible quadratics in $\mathbb{F}_3[x]$, namely,

$$p(x) = x^2 + 1, \quad q(x) = x^2 + x - 1, \quad \text{and} \quad r(x) = x^2 - x - 1;$$

each gives rise to a field with $9 = 3^2$ elements. Let us look at the first two in more detail. Proposition 3.117(v) says that $E = \mathbb{F}_3[x]/(p(x))$ is given by

$$E = \{a + b\alpha : \text{ where } \alpha^2 + 1 = 0\}.$$

Similarly, if $F = \mathbb{F}_3[x]/(q(x))$, then

$$F = \{a + b\beta : \text{ where } \beta^2 + \beta - 1 = 0\}.$$

These two fields are isomorphic, for the map $\varphi \colon E \to F$ (found by trial and error), defined by

$$\varphi(a + b\alpha) = a + b(1 - \beta),$$

is an isomorphism.

Now $\mathbb{F}_3[x]/(x^2 - x - 1)$ is also a field with nine elements, and it can shown that it is isomorphic to both of the two fields E and F just given (see Corollary 3.132).

(iii) In Example 3.35(ii) on page 137, we exhibited eight monic irreducible cubics $p(x) \in \mathbb{F}_3[x]$; each of them gives rise to a field $\mathbb{F}_3[x]/(p(x))$ having $27 = 3^3$ elements. ◂

We are now going to solve the isomorphism problem for finite fields.

Lemma 3.130. *Let $f(x) \in k[x]$, where k is a field, and let E be a splitting field of $f(x)$ over k. Let $\varphi \colon k \to k'$ be an isomorphism of fields, let $\varphi^* \colon k[x] \to k'[x]$ be the isomorphism*

$$g(x) = a_0 + a_1x + \cdots + a_nx^n \mapsto g^*(x) = \varphi(a_0) + \varphi(a_1)x + \cdots + \varphi(a_n)x^n,$$

and let E' be a splitting field of $f^*(x)$ over k'. Then there is an isomorphism $\Phi \colon E \to E'$ extending φ.

$$\begin{array}{ccc} E & \overset{\Phi}{\dashrightarrow} & E' \\ \big| & & \big| \\ k & \underset{\varphi}{\longrightarrow} & k' \end{array}$$

Proof. The proof is by induction on $d = [E : k]$. If $d = 1$, then $f(x)$ is a product of linear polynomials in $k[x]$, and it follows easily that $f^*(x)$ is also a product of linear polynomials in $k'[x]$. Therefore, $E' = k'$, and we may set $\Phi = \varphi$.

For the inductive step, choose a root z of $f(x)$ in E that is not in k, and let $p(x) = \operatorname{irr}(z, k)$ be the minimal polynomial of z over k (Proposition 3.117). Now $\deg(p) > 1$, because $z \notin k$; moreover, $[k(z) : k] = \deg(p)$, by Theorem 3.117. Let z' be a root of $p^*(x)$ in E', and let $p^*(x) = \operatorname{irr}(z', k')$ be the corresponding monic irreducible polynomial in $k'[x]$.

By a straightforward generalization[16] of Proposition 3.120(ii), there is an isomorphism $\widetilde{\varphi} \colon k(z) \to k'(z')$ extending φ with $\widetilde{\varphi} \colon z \mapsto z'$. We may regard $f(x)$ as a polynomial with

[16]Proving the generalization earlier would have involved introducing all the notation in the present hypothesis, and so it would have made a simple result appear complicated. The isomorphism $\varphi \colon k \to k'$ induces an isomorphism $\varphi^* \colon k[x] \to k'[x]$, which takes $p(x)$ to some polynomial $p^*(x)$, and φ^* induces an isomorphism $k[x]/(p(x)) \to k'[x]/(p^*(x))$.

coefficients in $k(z)$ (for $k \subseteq k(z)$ implies $k[x] \subseteq k(z)[x]$). We claim that E is a splitting field of $f(x)$ over $k(z)$; that is,

$$E = k(z)(z_1, \ldots, z_n),$$

where $z_1, \ldots, z_n$ are the roots of $f(x)/(x - z)$; after all,

$$E = k(z, z_1, \ldots, z_n) = k(z)(z_1, \ldots, z_n).$$

Similarly, E' is a splitting field of $f^*(x)$ over $k'(z')$. But $[E : k(z)] < [E : k]$, by Theorem 3.121, so that the inductive hypothesis gives an isomorphism $\Phi \colon E \to E'$ that extends $\widetilde{\varphi}$, and hence φ. •

Theorem 3.131. *If k is a field and $f(x) \in k[x]$, then any two splitting fields of $f(x)$ over k are isomorphic via an isomorphism that fixes k pointwise.*

Proof. Let E and E' be splitting fields of $f(x)$ over k. If φ is the identity, then the theorem applies at once. •

It is remarkable that the next theorem was not proved until the 1890s, 60 years after Galois discovered finite fields.

Corollary 3.132 (*E. H. Moore*). *Any two finite fields having exactly p^n elements are isomorphic.*

Proof. If E is a field with $q = p^n$ elements, then Lagrange's theorem applied to the multiplicative group $E^\times$ shows that $a^{q-1} = 1$ for every $a \in E^\times$. It follows that every element of E is a root of $f(x) = x^q - x \in \mathbb{F}_p[x]$, and so E is a splitting field of $f(x)$ over $\mathbb{F}_p$. •

E. H. Moore (1862–1932) began his mathematical career as an algebraist, but he did important work in many other parts of mathematics as well; for example, Moore–Smith convergence is named in part after him.

Finite fields are often called ***Galois fields*** in honor of their discoverer. In light of Corollary 3.132, we may speak of *the* field with q elements, where $q = p^n$ is a power of a prime p.

EXERCISES

3.81 Prove that if $I = \{0\}$, then $R/I \cong R$.

3.82 (***Third Isomorphism Theorem for Rings***) If R is a commutative ring having ideals $I \subseteq J$, then J/I is an ideal in R/I and there is a ring isomorphism $(R/I)/(J/I) \cong R/J$.

3.83 For every commutative ring R, prove that $R[x]/(x) \cong R$.

3.84 Prove that $\mathbb{F}_3[x]/(x^3 - x^2 + 1) \cong \mathbb{F}_3[x]/(x^3 - x^2 + x + 1)$.

3.85 If X is a subset of a commutative ring R, define $\mathcal{I}(X)$ to be the intersection of all those ideals I in R that contain X. Prove that $\mathcal{I}(X)$ is the set of all $a \in R$ for which there exist finitely many elements $x_1, \ldots, x_n \in X$ and elements $r_i \in R$ with $a = r_1 x_1 + \cdots + r_n x_n$.

3.86 Let $h(x), p(x) \in k[x]$ be monic polynomials, where k is a field. If $p(x)$ is irreducible and if every root of $h(x)$ (in an appropriate splitting field) is also a root of $p(x)$, prove that $h(x) = p(x)^m$ for some integer $m \geq 1$.
Hint. Use induction on $\deg(h)$.

3.87 ***Chinese Remainder Theorem***.

(i) Prove that if k is a field and $f(x), f'(x) \in k[x]$ are relatively prime, then given $b(x), b'(x) \in k[x]$, there exists $c(x) \in k[x]$ with

$$c - b \in (f) \text{ and } c - b' \in (f');$$

moreover, if $d(x)$ is another common solution, then $c - d \in (ff')$.

Hint. Adapt the proof of Theorem 1.28. This exercise is generalized to commutative rings in Exercise 6.11(iii) on page 325.

(ii) Prove that if k is a field and $f(x), g(x) \in k[x]$ are relatively prime, then

$$k[x]/(f(x)g(x)) \cong k[x]/(f(x)) \times k[x]/(g(x)).$$

Hint. See the proof of Theorem 2.81.

3.88 (i) Prove that a field K cannot have subfields k' and k'' with $k' \cong \mathbb{Q}$ and $k'' \cong \mathbb{F}_p$ for some prime p.

(ii) Prove that a field K cannot have subfields k' and k'' with $k' \cong \mathbb{F}_p$ and $k'' \cong \mathbb{F}_q$, where $p \neq q$ are primes.

3.89 Prove that the stochastic group $\Sigma(2, \mathbb{F}_4) \cong A_4$.
Hint. See Exercise 3.19 on page 125.

3.90 Let $f(x) = s_0 + s_1 x + \cdots + s_{n-1} x^{n-1} + x^n \in k[x]$, where k is a field, and suppose that $f(x) = (x - \alpha_1)(x - \alpha_2) \cdots (x - \alpha_n)$. Prove that $s_{n-1} = -(\alpha_1 + \alpha_2 + \cdots + \alpha_n)$ and that $s_0 = (-1)^n \alpha_1 \alpha_2 \cdots \alpha_n$. Conclude that the sum and product of all the roots of $f(x)$ lie in k.

3.91 Write addition and multiplication tables for the field $\mathbb{F}_8$ with eight elements.
Hint. Use an irreducible cubic over $\mathbb{F}_2$.

3.92 Let $k \subseteq K \subseteq E$ be fields. Prove that if E is a finite extension of k, then E is a finite extension of K and K is a finite extension of k.
Hint. Use Corollary 3.90(ii).

3.93 Let $k \subseteq F \subseteq K$ be a tower of fields, and let $z \in K$. Prove that if $k(z)/k$ is finite, then $[F(z) : F] \leq [k(z) : k]$. In particular, $[F(z) : F]$ is finite.
Hint. Use Proposition 3.117 to obtain an irreducible polynomial $p(x) \in k[x]$; the polynomial $p(x)$ may factor in $K[x]$.

3.94 (i) Is $\mathbb{F}_4$ a subfield of $\mathbb{F}_8$?

(ii) For any prime p, prove that if $\mathbb{F}_{p^n}$ is a subfield of $\mathbb{F}_{p^m}$, then $n \mid m$ (the converse is also true, as we shall see later).

Hint. View $\mathbb{F}_{p^m}$ as a vector space over $\mathbb{F}_{p^n}$.

3.95 Let K/k be a field extension. If $A \subseteq K$ and $u \in k(A)$, prove that there are $a_1, \ldots, a_n \in A$ with $u \in k(a_1, \ldots, a_n)$.

4

Fields

4.1 INSOLVABILITY OF THE QUINTIC

This chapter will discuss what is nowadays called *Galois theory* (it was originally called *theory of equations*), the interrelation between field extensions and certain groups associated to them, called *Galois groups*. This theory will enable us to prove the theorem of Abel–Ruffini as well as Galois's theorem describing precisely when the quadratic formula can be generalized to polynomials of higher degree. Another corollary of this theory is a proof of the fundamental theorem of algebra.

By Kronecker's theorem, Theorem 3.123, for each monic $f(x) \in k[x]$, where k is a field, there is a field K containing k and (not necessarily distinct) roots $z_1, \ldots, z_n$ with

$$f(x) = x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0 = (x - z_1) \cdots (x - z_n).$$

By induction on $n \geq 1$, we can easily generalize[1] Exercise 3.90 on page 197:

$$\begin{cases} a_{n-1} = -\sum_i z_i \\ a_{n-2} = \sum_{i<j} z_i z_j \\ a_{n-3} = -\sum_{i<j<k} z_i z_j z_k \\ \quad\vdots \\ a_0 = (-1)^n z_1 z_2 \cdots z_n. \end{cases} \tag{1}$$

[1]The coefficients a_i may be viewed as polynomials in $z_1, \ldots, z_n$; as such, they are called the ***elementary symmetric polynomials***, for they are unchanged if the z's are permuted.

Notice that $-a_{n-1}$ is the sum of the roots and that $\pm a_0$ is the product of the roots. Given the coefficients of $f(x)$, can we find its roots; that is, given the a's, can we solve the system (1) of n equations in n unknowns? If $n = 2$, the answer is yes: The quadratic formula works. If $n = 3$ or 4, the answer is still yes, for the cubic and quartic formulas work. But if $n \geq 5$, we shall see that no *analogous* solution exists.

We did not say that no solution of system (1) exists if $n \geq 5$; we said that no solution analogous to the solutions of the classical formulas exists. It is quite possible that there is some way of finding the roots of a quintic polynomial if we do not limit ourselves to field operations and extraction of roots only. Indeed, we can find the roots by *Newton's method*: if r is a real root of a polynomial $f(x)$ and if x_0 is a "good" approximation to r, then $r = \lim_{n\to\infty} x_n$, where x_n is defined recursively by $x_{n+1} = x_n - f(x_n)/f'(x_n)$ for all $n \geq 0$. There is a method of Hermite finding roots of quintics using elliptic modular functions, and there are methods for finding the roots of many polynomials of higher degree using hypergeometric functions.

We are going to show, if $n \geq 5$, that there is no solution "by radicals" (we will define this notion more carefully later). The key observation is that symmetry is present. Recall from Chapter 2 that if Ω is a polygon in the plane $\mathbb{R}^2$, then its symmetry group $\Sigma(\Omega)$ consists of all those motions $\varphi\colon \mathbb{R}^2 \to \mathbb{R}^2$ of the plane for which $\varphi(\Omega) = \Omega$. Moreover, motions $\varphi \in \Sigma(\Omega)$ are completely determined by their values on the vertices of Δ; indeed, if Ω has n vertices, then $\Sigma(\Omega)$ is isomorphic to a subgroup of S_n.

We are going to set up an analogy with symmetry groups in which polynomials play the role of polygons, a splitting field of a polynomial plays the role of the plane $\mathbb{R}^2$, and an *automorphism fixing k* plays the role of a motion.

Definition. Let E be a field containing a subfield k. An ***automorphism***[2] of E is an isomorphism $\sigma\colon E \to E$; we say that σ ***fixes*** k if $\sigma(a) = a$ for every $a \in k$.

For example, consider $f(x) = x^2 + 1 \in \mathbb{Q}[x]$. A splitting field of $f(x)$ over $\mathbb{Q}$ is $E = \mathbb{Q}(i)$, and complex conjugation $\sigma\colon a \mapsto \overline{a}$ is an example of an automorphism of E fixing $\mathbb{Q}$.

Proposition 4.1. *Let k be a subfield of a field K, let*

$$f(x) = x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0 \in k[x],$$

and let $E = k(z_1, \ldots, z_n) \subseteq K$ be a splitting field. If $\sigma\colon E \to E$ is an automorphism fixing k, then σ permutes the set of roots $\{z_1, \ldots, z_n\}$ of $f(x)$.

Proof. If r is a root of $f(x)$, then

$$0 = f(r) = r^n + a_{n-1}r^{n-1} + \cdots + a_1 r + a_0.$$

[2]The word *automorphism* is made up of two Greek roots: *auto*, meaning "self," and *morph*, meaning "shape" or "form." Just as an isomorphism carries one group onto an identical replica, an automorphism carries a group onto itself.

Applying σ to this equation gives

$$\begin{aligned} 0 &= \sigma(r)^n + \sigma(a_{n-1})\sigma(r)^{n-1} + \cdots + \sigma(a_1)\sigma(r) + \sigma(a_0) \\ &= \sigma(r)^n + a_{n-1}\sigma(r)^{n-1} + \cdots + a_1\sigma(r) + a_0 \\ &= f(\sigma(r)), \end{aligned}$$

because σ fixes k. Therefore, $\sigma(r)$ is a root of $f(x)$; thus, if Z is the set of all the roots, then $\sigma|Z \colon Z \to Z$, where $\sigma|Z$ is the restriction. But $\sigma|Z$ is injective (because σ is), so that Exercise 1.58 on page 36 says that $\sigma|Z$ is a permutation of Z. •

Here is the analog of the symmetry group $\Sigma(\Omega)$ of a polygon Ω.

Definition. Let k be a subfield of a field E. The ***Galois group*** of E over k, denoted by $\mathrm{Gal}(E/k)$, is the set of all those automorphisms of E that fix k. If $f(x) \in k[x]$, and if $E = k(z_1, \ldots, z_n)$ is a splitting field, then the ***Galois group*** of $f(x)$ over k is defined to be $\mathrm{Gal}(E/k)$.

It is easy to check that $\mathrm{Gal}(E/k)$ is a group with operation composition of functions. This definition is due to E. Artin (1898–1962), in keeping with his and E. Noether's emphasis on "abstract" algebra. Galois's original version (a group isomorphic to this one) was phrased, not in terms of automorphisms, but in terms of certain permutations of the roots of a polynomial (see Tignol, *Galois' Theory of Algebraic Equations*, pages 306–331). Note that $\mathrm{Gal}(E/k)$ is independent of the choice of splitting field E, by Theorem 3.131.

The following lemma will be used several times.

Lemma 4.2. *Let $E = k(z_1, \ldots, z_n)$. If $\sigma \colon E \to E$ is an automorphism fixing k, that is, if $\sigma \in \mathrm{Gal}(E/k)$, and if $\sigma(z_i) = z_i$ for all i, then σ is the identity 1_E.*

Proof. We prove the lemma by induction on $n \geq 1$. If $n = 1$, then each $u \in E$ has the form $u = f(z_1)/g(z_1)$, where $f(x), g(x) \in k[x]$ and $g(z_1) \neq 0$. But σ fixes z_1 as well as the coefficients of $f(x)$ and of $g(x)$, so that σ fixes all $u \in E$. For the inductive step, write $K = k(z_1, \ldots, z_{n-1})$, and note that $E = K(z_n)$ [for $K(z_n)$ is the smallest subfield containing k and $z_1, \ldots, z_{n-1}, z_n$]. Having noted this, the inductive step is just a repetition of the base step with k replaced by K. •

Theorem 4.3. *If $f(x) \in k[x]$ has degree n, then its Galois group $\mathrm{Gal}(E/k)$ is isomorphic to a subgroup of S_n.*

Proof. Let $X = \{z_1, \ldots, z_n\}$. If $\sigma \in \mathrm{Gal}(E/k)$, then Proposition 4.1 shows that its restriction $\sigma|X$ is a permutation of X; that is, $\sigma|X \in S_X$. Define $\varphi \colon \mathrm{Gal}(E/k) \to S_X$ by $\varphi \colon \sigma \mapsto \sigma|X$. To see that φ is a homomorphism, note that both $\varphi(\sigma\tau)$ and $\varphi(\sigma)\varphi(\tau)$ are functions $X \to X$, and hence they are equal if they agree on each $z_i \in X$. But $\varphi(\sigma\tau) \colon z_i \mapsto (\sigma\tau)(z_i)$, while $\varphi(\sigma)\varphi(\tau) \colon z_i \mapsto \sigma(\tau(z_i))$, and these are the same.

The image of φ is a subgroup of $S_X \cong S_n$. The kernel of φ is the set of all $\sigma \in \mathrm{Gal}(E/k)$ such that σ is the identity permutation on X; that is, σ fixes each of the roots z_i. As σ also fixes k, by definition of the Galois group, Lemma 4.2 gives $\ker\varphi = \{1\}$. Therefore, φ is injective, giving the theorem. •

If $f(x) = x^2 + 1 \in \mathbb{Q}[x]$, then complex conjugation σ is an automorphism of its splitting field $\mathbb{Q}(i)$ which fixes $\mathbb{Q}$ (and interchanges the roots i and $-i$). Since $\mathrm{Gal}(\mathbb{Q}(i)/\mathbb{Q})$ is a subgroup of the symmetric group S_2, which has order 2, it follows that $\mathrm{Gal}(\mathbb{Q}(i)/\mathbb{Q}) = \langle\sigma\rangle \cong \mathbb{I}_2$. We should regard the elements of any Galois group $\mathrm{Gal}(E/k)$ as generalizations of complex conjugation.

We are going to compute the order of the Galois group, but we first obtain some information about field isomorphisms and automorphisms.

Lemma 4.4. *If k is a field of characteristic 0, then every irreducible polynomial $p(x) \in k[x]$ has no repeated roots.*

Proof. In Exercise 3.37 on page 142, we saw, for any (not necessarily irreducible) polynomial $f(x)$ with coefficients in any field, that $f(x)$ has no repeated roots if and only if the gcd $(f, f') = 1$, where $f'(x)$ is the derivative of $f(x)$.

Now consider $p(x) \in k[x]$. Either $p'(x) = 0$ or $\deg(p') < \deg(p)$. Since $p(x)$ is irreducible, it is not constant, and so it has some nonzero monomial $a_i x^i$, where $i \geq 1$. Therefore, $i a_i x^{i-1}$ is a nonzero monomial in $p'(x)$, because k has characteristic 0, and so $p'(x) \neq 0$. Finally, since $p(x)$ is irreducible, its only divisors are constants and associates; as $p'(x)$ has smaller degree, it is not an associate of $p(x)$, and so the gcd $(p', p) = 1$. •

Recall Theorem 3.120(i): If E/k is an extension and $\alpha \in E$ is algebraic over k, then there is a unique monic irreducible polynomial $\mathrm{irr}(\alpha, k) \in k[x]$, called its *minimal polynomial*, having α as a root.

Definition. Let E/k be an algebraic extension. An *irreducible* polynomial $p(x)$ is ***separable*** if it has no repeated roots. An arbitrary polynomial $f(x)$ is ***separable*** if each of its irreducible factors has no repeated roots.

An element $\alpha \in E$ is called ***separable*** if either α is transcendental over k or if α is algebraic over k and its minimal polynomial $\mathrm{irr}(\alpha, k)$ has no repeated roots; that is, $\mathrm{irr}(\alpha, k)$ is a separable polynomial.

A field extension E/k is called a ***separable extension*** if each of its elements is separable; E/k is ***inseparable*** if it is not separable.

Lemma 4.4 shows that every extension of a field of characteristic 0 is a separable extension. If E is a finite field with p^n elements, then Lagrange's theorem (for the multiplicative group $E^\times$) shows that every element of E is a root of $x^{p^n} - x$. We saw, in the proof of Theorem 3.127 (the existence of finite fields with p^n elements), that $x^{p^n} - x$ has no repeated roots. It follows that if $k \subseteq E$, then E/k is a separable extension, for if $\alpha \in E$, then $\mathrm{irr}(\alpha, k)$ is a divisor of $x^{p^n} - x$.

Example 4.5.

Here is an example of an inseparable extension. Let $k = \mathbb{F}_p(t) = \mathrm{Frac}(\mathbb{F}_p[t])$, and let $E = k(\alpha)$, where α is a root of $f(x) = x^p - t$; that is, $\alpha^p = t$. In $E[x]$, we have

$$f(x) = x^p - t = x^p - \alpha^p = (x - \alpha)^p.$$

If we show that $\alpha \notin k$, then $f(x)$ is irreducible, by Proposition 3.126, and so $f(x) = \operatorname{irr}(\alpha, k)$ is an inseparable polynomial. Therefore, E/k is an inseparable extension.

It remains to show that $\alpha \notin k$. Otherwise, there are $g(t), h(t) \in \mathbb{F}_p[t]$ with $\alpha = g(t)/h(t)$. Hence, $g = \alpha h$ and $g^p = \alpha^p h^p = th^p$, so that

$$\deg(g^p) = \deg(th^p) = 1 + \deg(h^p).$$

But $p \mid \deg(g^p)$ and $p \mid \deg(h^p)$, and this gives a contradiction. ◀

We will study separability and inseparability more thoroughly in Chapter 6.

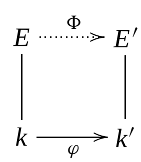
Example 4.6.
Let m be a positive integer, let k be a field, and let $f(x) = x^m - 1 \in k[x]$. If the characteristic of k does not divide m, then $mx^{m-1} \neq 0$ and the gcd $(f, f') = 1$; hence, $f(x)$ has no repeated roots. Therefore, any splitting field E/k of $f(x)$ contains m distinct mth roots of unity. Moreover, the set of these roots of unity is a (multiplicative) subgroup of $E^\times$ of order m that is cyclic, by Theorem 3.30. We have proved that if characteristic $k \nmid m$, then there exists a primitive mth root of unity ω in some extension field of k, and ω is a separable element.

On the other hand, if p^e is a prime power and k has characteristic p, then $x^{p^e} - 1 = (x-1)^{p^e}$, and so there is only one p^eth root of unity, namely, 1. ◀

Separability of E/k allows us to find the order of $\operatorname{Gal}(E/k)$.

Theorem 4.7.

(i) *Let E/k be a splitting field of a separable polynomial $f(x) \in k[x]$, let $\varphi\colon k \to k'$ be a field isomorphism, and let E'/k' be a splitting field of $f^*(x) \in k'[x]$ [where $f^*(x)$ is obtained from $f(x)$ by applying φ to its coefficients].*



Then there are exactly $[E : k]$ isomorphisms $\Phi\colon E \to E'$ that extend φ.

(ii) *If E/k is a splitting field of a separable $f(x) \in k[x]$, then*

$$|\operatorname{Gal}(E/k)| = [E : k].$$

Proof. (i) The proof, by induction on $[E : k]$, modifies that of Lemma 3.130. If $[E : k] = 1$, then $E = k$ and there is only one extension Φ of φ, namely, φ itself. If $[E : k] > 1$, let $f(x) = p(x)g(x)$, where $p(x)$ is an irreducible factor of largest degree, say, d. We may assume that $d > 1$, otherwise $f(x)$ splits over k and $[E : k] = 1$. Choose a root α of $p(x)$ (note that $\alpha \in E$ because E is a splitting field of $f(x) = p(x)g(x)$). If $\widetilde{\varphi}\colon k(\alpha) \to E'$

is any extension of φ, then $\varphi(\alpha)$ is a root α' of $p^*(x)$, by Proposition 4.1; since $f^*(x)$ is separable, $p^*(x)$ has exactly d roots $\alpha' \in E'$; by Lemma 4.2 and Theorem 3.120(ii), there are exactly d isomorphisms $\widehat{\varphi} : k(\alpha) \to k'(\alpha')$ extending φ, one for each α'. Now E is also a splitting field of $f(x)$ over $k(\alpha)$, because adjoining all the roots of $f(x)$ to $k(\alpha)$ still produces E, and E' is a splitting field of $f^*(x)$ over $k'(\alpha')$. Since $[E : k(\alpha)] = [E : k]/d$, induction shows that each of the d isomorphisms $\widehat{\varphi}$ has exactly $[E : k]/d$ extensions $\Phi\colon E \to E'$. Thus, we have constructed $[E : k]$ isomorphisms extending φ. But there are no others, because every τ extending φ has $\tau|k(\alpha) = \widehat{\varphi}$ for some $\widehat{\varphi}\colon k(\alpha) \to k'(\alpha')$.

(ii) In part (i), take $k = k'$, $E = E'$, and $\varphi = 1_k$. •

Example 4.8.
The separability hypothesis in Theorem 4.7(ii) is necessary. In Example 4.5, we saw that if $k = \mathbb{F}_p(t)$ and α is a root of $x^p - t$, then $E = k(\alpha)$ is an inseparable extension. Moreover, $x^p - t = (x - \alpha)^p$, so that α is the only root of this polynomial. Therefore, if $\sigma \in \mathrm{Gal}(E/k)$, then Proposition 4.1 shows that $\sigma(\alpha) = \alpha$. Therefore, $\mathrm{Gal}(E/k) = \{1\}$, by Lemma 4.2, and so $|\mathrm{Gal}(E/k)| < [E : k] = p$ in this case. ◄

Corollary 4.9. *Let E/k be a splitting field of a separable polynomial $f(x) \in k[x]$ of degree n. If $f(x)$ is irreducible, then $n \mid |\mathrm{Gal}(E/k)|$.*

Proof. By the theorem, $|\mathrm{Gal}(E/k)| = [E : k]$. Let $\alpha \in E$ be a root of $f(x)$. Since $f(x)$ is irreducible, $[k(\alpha) : k] = n$, and

$$[E : k] = [E : k(\alpha)][k(\alpha) : k] = n[E : k(\alpha)]. \quad \bullet$$

We shall see, in Proposition 4.38, that if E/k is a splitting field of a separable polynomial, then E/k is a separable extension.

Here are some computations of Galois groups of specific polynomials in $\mathbb{Q}[x]$.

Example 4.10.
(i) Let $f(x) = x^3 - 1 \in \mathbb{Q}[x]$. Now $f(x) = (x - 1)(x^2 + x + 1)$, where $x^2 + x + 1$ is irreducible (the quadratic formula shows that its roots ω and $\overline{\omega}$, do not lie in $\mathbb{Q}$). The splitting field of $f(x)$ is $\mathbb{Q}(\omega)$, for $\omega^2 = \overline{\omega}$, and so $[\mathbb{Q}(\omega) : \mathbb{Q}] = 2$. Therefore, $|\mathrm{Gal}(\mathbb{Q}(\omega)/\mathbb{Q})| = 2$, by Theorem 4.7(ii), and it is cyclic of order 2. Its nontrivial element is complex conjugation.

(ii) Let $f(x) = x^2 - 2 \in \mathbb{Q}[x]$. Now $f(x)$ is irreducible with roots $\pm\sqrt{2}$, so that $E = \mathbb{Q}(\sqrt{2})$ is a splitting field. By Theorem 4.7(ii), $|\mathrm{Gal}(E/\mathbb{Q})| = 2$. Now every element of E has a unique expression of the form $a + b\sqrt{2}$, where $a, b \in \mathbb{Q}$ [Theorem 3.117(v)], and it is easily seen that $\sigma\colon E \to E$, defined by $\sigma\colon a + b\sqrt{2} \mapsto a - b\sqrt{2}$, is an automorphism of E fixing $\mathbb{Q}$. Therefore, $\mathrm{Gal}(E/\mathbb{Q}) = \langle\sigma\rangle$, where σ interchanges $\sqrt{2}$ and $-\sqrt{2}$.

(iii) Let $g(x) = x^3 - 2 \in \mathbb{Q}[x]$. The roots of $g(x)$ are α, $\omega\alpha$, and $\omega^2\alpha$, where $\alpha = \sqrt[3]{2}$, the real cube root of 2, and ω is a primitive cube root of unity. It is easy to see that the splitting field of $g(x)$ is $E = \mathbb{Q}(\alpha, \omega)$. Note that

$$[E : \mathbb{Q}] = [E : \mathbb{Q}(\alpha)][\mathbb{Q}(\alpha) : \mathbb{Q}] = 3[E : \mathbb{Q}(\alpha)],$$

for $g(x)$ is irreducible over $\mathbb{Q}$ (it is a cubic having no rational roots). Now $E \neq \mathbb{Q}(\alpha)$, for every element in $\mathbb{Q}(\alpha)$ is real, while the complex number ω is not real. Therefore, $[E : \mathbb{Q}] = |\operatorname{Gal}(E/\mathbb{Q})| > 3$. On the other hand, we know that $\operatorname{Gal}(E/\mathbb{Q})$ is isomorphic to a subgroup of S_3, and so we must have $\operatorname{Gal}(E/\mathbb{Q}) \cong S_3$.

(iv) We examined $f(x) = x^4 - 10x^2 + 1 \in \mathbb{Q}[x]$ in Example 3.122, when we saw that $f(x)$ is irreducible; in fact, $f(x) = \operatorname{irr}(\beta, \mathbb{Q})$, where $\beta = \sqrt{2} + \sqrt{3}$. If $E = \mathbb{Q}(\beta)$, then $[E : \mathbb{Q}] = 4$; moreover, E is a splitting field of $f(x)$, where the other roots of $f(x)$ are $-\sqrt{2} - \sqrt{3}$, $-\sqrt{2} + \sqrt{3}$, and $\sqrt{2} - \sqrt{3}$. It follows from Theorem 4.7(ii) that if $G = \operatorname{Gal}(E/\mathbb{Q})$, then $|G| = 4$; hence, either $G \cong \mathbb{I}_4$ or $G \cong \mathbf{V}$.

We also saw, in Example 3.122, that E contains $\sqrt{2}$ and $\sqrt{3}$. If σ is an automorphism of E fixing $\mathbb{Q}$, then $\sigma(\sqrt{2}) = u\sqrt{2}$, where $u = \pm 1$, because $(\sigma(\sqrt{2}))^2 = 2$. Therefore, $\sigma^2(\sqrt{2}) = \sigma(u\sqrt{2}) = u\sigma(\sqrt{2}) = u^2\sqrt{2} = \sqrt{2}$; similarly, $\sigma^2(\sqrt{3}) = \sqrt{3}$. If α is a root of $f(x)$, then $\alpha = u\sqrt{2} + v\sqrt{3}$, where $u, v = \pm 1$. Hence,

$$\sigma^2(\alpha) = u\sigma^2(\sqrt{2}) + v\sigma^2(\sqrt{3}) = u\sqrt{2} + v\sqrt{3} = \alpha.$$

Lemma 4.2 gives $\sigma^2 = 1_E$ for all $\sigma \in \operatorname{Gal}(E/\mathbb{Q})$, and so $\operatorname{Gal}(E/\mathbb{Q}) \cong \mathbf{V}$.

Here is another way to compute $G = \operatorname{Gal}(E/\mathbb{Q})$. We saw in Example 3.122 that $E = \mathbb{Q}(\sqrt{2}+\sqrt{3}) = \mathbb{Q}(\sqrt{2}, \sqrt{3})$ is also a splitting field of $g(x) = (x^2-2)(x^2-3)$ over $\mathbb{Q}$. By Proposition 3.120(ii), there is an automorphism $\varphi\colon \mathbb{Q}(\sqrt{2}) \to \mathbb{Q}(\sqrt{2})$ taking $\sqrt{2} \mapsto -\sqrt{2}$. But $\sqrt{3} \notin \mathbb{Q}(\sqrt{2})$, as we noted in Example 3.122, so that $x^2 - 3$ is irreducible over $\mathbb{Q}(\sqrt{2})$. Lemma 3.130 shows that φ extends to an automorphism $\Phi\colon E \to E$; of course, $\Phi \in \operatorname{Gal}(E/\mathbb{Q})$. There are two possibilites: $\Phi(\sqrt{3}) = \pm\sqrt{3}$. Indeed, it is now easy to see that the elements of $\operatorname{Gal}(E/\mathbb{Q})$ correspond to the four-group, consisting of the identity and the permutations (in cycle notation)

$$\big(\sqrt{2},\ -\sqrt{2}\big)\big(\sqrt{3},\ \sqrt{3}\big), \quad \big(\sqrt{2},\ -\sqrt{2}\big)\big(\sqrt{3},\ -\sqrt{3}\big), \quad \big(\sqrt{2},\ \sqrt{2}\big)\big(\sqrt{3},\ -\sqrt{3}\big). \quad ◀$$

Here are two more general computations of Galois groups.

Proposition 4.11. *If m is a positive integer, if k is a field, and if E is a splitting field of $x^m - 1$ over k, then $\operatorname{Gal}(E/k)$ is abelian; in fact, $\operatorname{Gal}(E/k)$ is isomorphic to a subgroup of the multiplicative group $U(\mathbb{I}_m)$ of all $[i]$ with $(i, m) = 1$.*

Proof. Assume first that the characteristic of k does not divide m. By Example 4.6, E contains a primitive mth root of unity, ω, and so $E = k(\omega)$. The group of all roots of $x^m - 1$ in E is cyclic, say, with generator ω, so that if $\sigma \in \operatorname{Gal}(E/k)$, then its restriction is an automorphism of the cyclic group $\langle\omega\rangle$. Hence, $\sigma(\omega) = \omega^i$ must also be a generator of $\langle\omega\rangle$; that is, $(i, m) = 1$, by Theorem 2.33(i). It is easy to see that i is uniquely determined mod m, so that the function $\varphi\colon \operatorname{Gal}(k(\omega)/k) \to U(\mathbb{I}_m)$, given by $\varphi(\sigma) = [i]$ if $\sigma(\omega) = \omega^i$, is well-defined. Now φ is a homomorphism, for if $\tau(\omega) = \omega^j$, then

$$\tau\sigma(\omega) = \tau(\omega^i) = (\omega^i)^j = \omega^{ij}.$$

Finally, Lemma 4.2 shows that φ is injective.

Suppose now that k has characteristic p and that $m = p^e n$, where $p \nmid n$. By Example 4.6, there is a primitive nth root of unity ω, and we claim that $E = k(\omega)$ is a splitting field of $x^m - 1$. If $\zeta^m = 1$, then $1 = \zeta^{p^e n} = (\zeta^n)^{p^e}$. But the only p^eth root of unity is 1, since k has characteristic p, and so $\zeta^n = 1$; that is, $\zeta \in k(\omega)$. We have reduced to the case of the first paragraph. [In fact, more is true in this case: $\mathrm{Gal}(E/k)$ is isomorphic to a subgroup of the multiplicative group $U(\mathbb{I}_n)$.] •

Remark. We cannot conclude more from the proposition; given any finite abelian group G, there is some integer m with G isomorphic to a subgroup of $U(\mathbb{I}_m)$. ◂

Theorem 4.12. *If p is a prime, then*

$$\mathrm{Gal}(\mathbb{F}_{p^n}/\mathbb{F}_p) \cong \mathbb{I}_n,$$

and a generator is the ***Frobenius*** $F : u \mapsto u^p$.

Proof. Let $q = p^n$, and let $G = \mathrm{Gal}(\mathbb{F}_q/\mathbb{F}_p)$. Since $\mathbb{F}_q$ has characteristic p, we have $(a + b)^p = a^p + b^p$, and so the Frobenius F is a homomorphism of fields. As any homomorphism of fields, F is injective; as $\mathbb{F}_q$ is finite, F must be an automorphism, by Exercise 1.58 on page 36; that is, $F \in G$.

If $\pi \in \mathbb{F}_q$ is a primitive element, then $d(x) = \mathrm{irr}(\pi, \mathbb{F}_p)$ has degree n, by Corollary 3.128, and so $|G| = n$, by Theorem 4.7(ii). It suffices to prove that the order j of F is not less than n. But if $F^j = 1_{\mathbb{F}_q}$ for $j < n$, then $u^{p^j} = u$ for all of the $q = p^n$ elements $u \in \mathbb{F}_q$, giving too many roots of the polynomial $x^{p^j} - x$. •

The following nice corollary of Lemma 3.130 says, in our analogy between Galois theory and symmetry of polygons, that irreducible polynomials correspond to regular polygons.

Proposition 4.13. *Let k be a field and let $p(x) \in k[x]$ have no repeated roots. If E/k is a splitting field of $p(x)$, then $p(x)$ is irreducible if and only if* $\mathrm{Gal}(E/k)$ *acts transitively on the roots of $p(x)$.*

Proof. Assume that $p(x)$ is irreducible, and let $\alpha, \beta \in E$ be roots of $p(x)$. By Theorem 3.120(i), there is an isomorphism $\varphi : k(\alpha) \to k(\beta)$ with $\varphi(\alpha) = \beta$ and which fixes k. Lemma 3.130 shows that φ extends to an automorphism Φ of E that fixes k; that is, $\Phi \in \mathrm{Gal}(E/k)$. Now $\Phi(\alpha) = \varphi(\alpha) = \beta$, and so $\mathrm{Gal}(E/k)$ acts transitively on the roots.

Conversely, assume that $\mathrm{Gal}(E/k)$ acts transitively on the roots of $p(x)$. If $p(x) = q_1(x) \cdots q_t(x)$ is a factorization into irreducibles in $k[x]$, where $t \geq 2$, choose a root $\alpha \in E$ of $q_1(x)$ and choose a root $\beta \in E$ of $q_2(x)$. By hypothesis, there is $\sigma \in \mathrm{Gal}(E/k)$ with $\sigma(\alpha) = \beta$. Now σ permutes the roots of $q_1(x)$, by Proposition 4.1. However, β is not a root of $q_1(x)$, because $p(x)$ has no repeated roots, and this is a contradiction. Therefore, $t = 1$; that is, $p(x)$ is irreducible. •

We can now give another proof of Corollary 4.9. Theorem 2.98 says that if X is a G-set, then $|G| = |\mathcal{O}(x)||G_x|$, where $\mathcal{O}(x)$ is the orbit of $x \in X$. In particular, if X is a transitive G-set, then $|X|$ is a divisor of $|G|$. Let $f(x) \in k[x]$ be a separable irreducible polynomial of degree n, and let E/k be its splitting field. If X is the set of roots of $f(x)$, then X is a transitive $\mathrm{Gal}(E/k)$-set, by Proposition 4.13, and so $n = \deg(f) = |X|$ is a divisor of $|\mathrm{Gal}(E/k)|$.

The analogy[3] is complete.

Polygon Ω	polynomial $f(x) \in k[x]$
Regular polygon	irreducible polynomial
Vertices of Ω	roots of $f(x)$
Plane	splitting field E of $f(x)$
Motion	automorphism fixing k
Symmetry group $\Sigma(\Omega)$	Galois group $\mathrm{Gal}(E/k)$

Here is the basic strategy. First, we will translate the *classical formulas* (giving the roots of polynomials of degree at most 4) in terms of subfields of a splitting field E over k. Second, this translation into the language of fields will itself be translated into the language of groups: If there is a formula for the roots of $f(x)$, then $\mathrm{Gal}(E/k)$ must be a *solvable* group (which we will soon define). Finally, polynomials of degree at least 5 can have Galois groups that are not solvable. The conclusion is that there are polynomials of degree 5 for which there is no formula, analogous to the classical formulas, giving their roots.

Formulas and Solvability by Radicals

Without further ado, here is the translation of the existence of a formula for the roots of a polynomial in terms of subfields of a splitting field.

Definition. A ***pure extension*** of ***type*** m is an extension $k(u)/k$, where $u^m \in k$ for some $m \geq 1$. An extension K/k is a ***radical extension*** if there is a tower of fields

$$k = K_0 \subseteq K_1 \subseteq \cdots \subseteq K_t = K$$

in which each K_{i+1}/K_i is a pure extension.

If $u^m = a \in k$, then $k(u)$ arises from k by adjoining an mth root of a. If $k \subseteq \mathbb{C}$, there are m different mth roots of a, namely, $u, \omega u, \omega^2 u, \ldots, \omega^{m-1}u$, where $\omega = e^{2\pi i/m}$ is a primitive mth root of unity. More generally, if k contains the mth roots of unity, then a pure extension $k(u)$ of type m, that is, $u^m = a \in k$, then $k(u)$ is a splitting field of $x^m - a$. Not every subfield k of $\mathbb{C}$ contains all the roots of unity; for example, 1 and -1 are the only roots of unity in $\mathbb{Q}$. Since we seek formulas involving extraction of roots, it will eventually be convenient to assume that k contains appropriate roots of unity.

[3]Actually, a better analogy would involve polyhedra in euclidean space $\mathbb{R}^n$ instead of only polygons in the plane.

When we say that there is a *formula* for the roots of a polynomial $f(x)$ analogous to the quadratic formula, we mean that there is some expression giving the roots of $f(x)$ in terms of the coefficients of $f(x)$. The expression may involve the field operations, constants, and extraction of roots, but it should not involve any other operations involving cosines, definite integrals, or limits, for example. We maintain that a formula as we informally described exists precisely when $f(x)$ is *solvable by radicals*, which we now define.

Definition. Let $f(x) \in k[x]$ have a splitting field E. We say that $f(x)$ is ***solvable by radicals*** if there is a radical extension

$$k = K_0 \subseteq K_1 \subseteq \cdots \subseteq K_t$$

with $E \subseteq K_t$.

Actually, there is a nontrivial result of Gauss that we are assuming. It is true, but not obvious, that $x^n - 1$ is solvable by radicals in the sense that there is the desired sort of expression for

$$e^{2\pi i/n} = \cos\left(\tfrac{2\pi}{n}\right) + i \sin\left(\tfrac{2\pi}{n}\right)$$

(see van der Waerden, *Modern Algebra* I, pages 163–168, or Tignol, *Galois' Theory of Algebraic Equations*, pages 252–256). This theorem of Gauss is what enabled him to construct a regular 17-gon with ruler and compass.

Let us illustrate this definition by considering the classical formulas for polynomials of small degree.

Quadratics

If $f(x) = x^2 + bx + c$, then the quadratic formula gives its roots as

$$\tfrac{1}{2}\left(-b \pm \sqrt{b^2 - 4c}\right).$$

Let $k = \mathbb{Q}(b, c)$. Define $K_1 = k(u)$, where $u = \sqrt{b^2 - 4c}$. Then K_1 is a radical extension of k, for $u^2 \in k$. Moreover, the quadratic formula implies that K_1 is the splitting field of $f(x)$, and so $f(x)$ is solvable by radicals.

Cubics

Let $f(X) = X^3 + bX^2 + cX + d$, and let $k = \mathbb{Q}(b, c, d)$. The change of variable $X = x - \frac{1}{3}b$ yields a new polynomial $\widetilde{f}(x) = x^3 + qx + r \in k[x]$ having the same splitting field E [for if u is a root of $\widetilde{f}(x)$, then $u - \frac{1}{3}b$ is a root of $f(x)$]; it follows that $\widetilde{f}(x)$ is solvable by radicals if and only if $f(x)$ is. Special cases of the cubic formula were discovered by Scipio del Ferro around 1515, and the remaining cases were completed by Niccolò Fontana (Tartaglia) in 1535 and by Giralamo Cardano in 1545. The formula gives the roots of $\widetilde{f}(x)$ as

$$g + h, \quad \omega g + \omega^2 h, \quad \text{and} \quad \omega^2 g + \omega h,$$

where $g^3 = \frac{1}{2}\left(-r + \sqrt{R}\right)$, $h = -q/3g$, $R = r^2 + \frac{4}{27}q^3$, and $\omega = -\frac{1}{2} + i\frac{\sqrt{3}}{2}$ is a primitive cube root of unity.

The cubic formula is derived as follows. If u is a root of $\widetilde{f}(x) = x^3 + qx + r$, write

$$u = g + h,$$

and substitute:

$$0 = \widetilde{f}(u) = \widetilde{f}(g + h) = g^3 + h^3 + (3gh + q)u + r.$$

Now the quadratic formula can be rephrased to say, given any pair of numbers u and v, that there are (possibly complex) numbers g and h with $u = g + h$ and $v = gh$. Therefore, we can further assume that $3gh + q = 0$; that is,

$$g^3 + h^3 = -r \qquad \text{and} \qquad gh = -\tfrac{1}{3}q.$$

After cubing the latter, the resulting pair of equations is

$$\begin{aligned} g^3 + h^3 &= -r \\ g^3h^3 &= -\tfrac{1}{27}q^3, \end{aligned}$$

giving the quadratic in g^3:

$$g^6 + rg^3 - \tfrac{1}{27}q^3 = 0.$$

The quadratic formula gives

$$g^3 = \tfrac{1}{2}\left(-r + \sqrt{r^2 + \tfrac{4}{27}q^3}\right) = \tfrac{1}{2}\left(-r + \sqrt{R}\right)$$

[note that h^3 is also a root of this quadratic, so that $h^3 = \frac{1}{2}\left(-r - \sqrt{R}\right)$]. There are three cube roots of g^3: g, ωg, and $\omega^2 g$. Because of the constraint $gh = -\frac{1}{3}q$, each of these has a "mate," namely, $h = -q/(3g)$, $-q/(3\omega g) = \omega^2 h$, and $-q/(3\omega^2 g) = \omega h$.

Let us now see that $\widetilde{f}(x)$ is solvable by radicals. Define $K_1 = k(\sqrt{R})$, where $R = r^2 + \frac{4}{27}q^3$, and define $K_2 = K_1(\alpha)$, where $\alpha^3 = \frac{1}{2}(-r + \sqrt{R})$. The cubic formula shows that K_2 contains the root $\alpha + \beta$ of $\widetilde{f}(x)$, where $\beta = -q/3\alpha$. Finally, define $K_3 = K_2(\omega)$, where $\omega^3 = 1$. The other roots of $\widetilde{f}(x)$ are $\omega\alpha + \omega^2\beta$ and $\omega^2\alpha + \omega\beta$, both of which lie in K_3, and so $E \subseteq K_3$.

A splitting field E need not equal K_3, for if all the roots of $f(x)$ are real, then $E \subseteq \mathbb{R}$, whereas $K_3 \not\subseteq \mathbb{R}$. An interesting aspect of the cubic formula is the so-called *casus irreducibilis*; the formula for the roots of an irreducible cubic in $\mathbb{Q}[x]$ having all roots real requires the presence of complex numbers (see Rotman, *Galois Theory*, 2d ed., page 99).

Casus Irreducibilis. If $f(x) = x^3 + qx + r \in \mathbb{Q}[x]$ is an irreducible polynomial having three real roots, then any radical extension $K_t/\mathbb{Q}$ containing the splitting field of $f(x)$ is not real; that is, $K_t \not\subseteq \mathbb{R}$.

Example 4.14.
If $f(x) = x^3 - 15x - 126$, then $q = -15$, $r = -126$, $R = 15376$, and $\sqrt{R} = 124$. Hence, $g^3 = 125$, so that $g = 5$. Thus, $h = -q/(3g) = 1$. Therefore, the roots of $f(x)$ are

$$6, \quad 5\omega + \omega^2 = -3 + 2i\sqrt{3}, \quad 5\omega^2 + \omega = -3 - 2i\sqrt{3}.$$

Alternatively, having found one root to be 6, the other two roots can be found as the roots of the quadratic $f(x)/(x-6) = x^2 + 6x + 21$. ◂

Example 4.15.
The cubic formula is not very useful because it often gives the roots in unrecognizable form. For example, let

$$f(x) = (x-1)(x-2)(x+3) = x^3 - 7x + 6.$$

The cubic formula gives

$$g + h = \sqrt[3]{\tfrac{1}{2}\Big(-6 + \sqrt{\tfrac{-400}{27}}\Big)} + \sqrt[3]{\tfrac{1}{2}\Big(-6 - \sqrt{\tfrac{-400}{27}}\Big)}.$$

It is not at all obvious that $g + h$ is a real number, let alone an integer. There is another version of the cubic formula, due to F. Viète, which gives the roots in terms of trigonometric functions instead of radicals (see my book, *A First Course in Abstract Algebra*, pp. 360–362). ◂

Quartics

Let $f(X) = X^4 + bX^3 + cX^2 + dX + e$, and let $k = \mathbb{Q}(b, c, d, e)$. The change of variable $X = x - \frac{1}{4}b$ yields a new polynomial $\widetilde{f}(x) = x^4 + qx^2 + rx + s \in k[x]$; moreover, the splitting field E of $f(x)$ is equal to the splitting field of $\widetilde{f}(x)$, for if u is a root of $\widetilde{f}(x)$, then $u - \frac{1}{4}b$ is a root of $f(x)$. The quartic formula was found by Luigi Ferrari in 1545, but here is the version presented by R. Descartes in 1637. Factor $\widetilde{f}(x)$ in $\mathbb{C}[x]$:

$$\widetilde{f}(x) = x^4 + qx^2 + rx + s = (x^2 + jx + \ell)(x^2 - jx + m),$$

and determine j, ℓ and m. Expanding and equating like coefficients gives the equations

$$\begin{aligned} \ell + m - j^2 &= q; \\ j(m - \ell) &= r; \\ \ell m &= s. \end{aligned}$$

The first two equations give

$$\begin{aligned} 2m &= j^2 + q + r/j; \\ 2\ell &= j^2 + q - r/j. \end{aligned}$$

Substituting these values for m and ℓ into the third equation yields the ***resolvent cubic***:

$$(j^2)^3 + 2q(j^2)^2 + (q^2 - 4s)j^2 - r^2.$$

The cubic formula gives j^2, from which we can determine m and ℓ, and hence the roots of the quartic.

Define pure extensions

$$k = K_0 \subseteq K_1 \subseteq K_2 \subseteq K_3,$$

as in the cubic case, so that $j^2 \in K_3$. Define $K_4 = K_3(j)$ (so that $\ell, m \in K_4$). Finally, define $K_5 = K_4\left(\sqrt{j^2 - 4\ell}\right)$ and $K_6 = K_5\left(\sqrt{j^2 - 4m}\right)$ [giving roots of the quadratic factors $x^2 + jx + \ell$ and $x^2 - jx + m$ of $\widetilde{f}(x)$]. The quartic formula gives $E \subseteq K_6$.

We have just seen that quadratics, cubics, and quartics are solvable by radicals. Conversely, if $f(x)$ is a polynomial that is solvable by radicals, then there is a formula of the desired kind that expresses its roots in terms of its coefficients. For suppose that

$$k = K_0 \subseteq K_1 \subseteq \cdots \subseteq K_t$$

is a radical extension with splitting field $E \subseteq K_t$. Let z be a root of $f(x)$. Now $K_t = K_{t-1}(u)$, where u is an mth root of some element $\alpha \in K_{t-1}$; hence, z can be expressed in terms of u and K_{t-1}; that is, z can be expressed in terms of $\sqrt[m]{\alpha}$ and K_{t-1}. But $K_{t-1} = K_{t-2}(v)$, where some power of v lies in K_{t-2}. Hence, z can be expressed in terms of u, v, and K_{t-2}. Ultimately, z is expressed by a formula analogous to those of the classical formulas.

Translation into Group Theory

The second stage of the strategy involves investigating the effect of $f(x)$ being solvable by radicals on its Galois group.

Suppose that $k(u)/k$ is a pure extension of type 6; that is, $u^6 \in k$. Now $k(u^3)/k$ is a pure extension of type 2, for $(u^3)^2 = u^6 \in k$, and $k(u)/k(u^3)$ is obviously a pure extension of type 3. Thus, $k(u)/k$ can be replaced by a tower of pure extensions $k \subseteq k(u^3) \subseteq k(u)$ of types 2 and 3. More generally, we may assume, given a tower of pure extensions, that each field is of prime type over its predecessor: If $k \subseteq k(u)$ is of type m, then factor $m = p_1 \cdots p_q$, where the p's are (not necessarily distinct) primes, and replace $k \subseteq k(u)$ by

$$k \subseteq k(u^{m/p_1}) \subseteq k(u^{m/p_1 p_2}) \subseteq \cdots \subseteq k(u).$$

Here is a key result allowing us to translate solvability by radicals into the language of Galois groups.

Theorem 4.16. *Let $k \subseteq B \subseteq E$ be a tower of fields, let $f(x), g(x) \in k[x]$, let B be a splitting field of $f(x)$ over k, and let E be a splitting field of $g(x)$ over k. Then* $\mathrm{Gal}(E/B)$ *is a normal subgroup of* $\mathrm{Gal}(E/k)$, *and*

$$\mathrm{Gal}(E/k)/\mathrm{Gal}(E/B) \cong \mathrm{Gal}(B/k).$$

Proof. Let $B = k(z_1, \ldots, z_t)$, where $z_1, \ldots, z_t$ are the roots of $f(x)$ in E. If $\sigma \in \mathrm{Gal}(E/k)$, then σ permutes $z_1, \ldots, z_t$, by Proposition 4.1(i) (for σ fixes k), and so $\sigma(B) = B$. Define $\rho\colon \mathrm{Gal}(E/k) \to \mathrm{Gal}(B/k)$ by $\sigma \mapsto \sigma|B$. It is easy to see, as in the proof of Theorem 4.3, that ρ is a homomorphism and that $\ker\rho = \mathrm{Gal}(E/B)$. It follows that $\mathrm{Gal}(E/B)$ is a normal subgroup of $\mathrm{Gal}(E/k)$. But ρ is surjective: If $\tau \in \mathrm{Gal}(B/k)$, then Lemma 3.130 applies to show that there is $\sigma \in \mathrm{Gal}(E/k)$ extending τ [i.e., $\rho(\sigma) = \sigma|B = \tau$]. The first isomorphism theorem completes the proof. •

The next technical result will be needed when we apply Theorem 4.16.

Lemma 4.17.

(i) *If $B = k(\alpha_1, \ldots, \alpha_n)$ is a finite extension of a field k, then there is a finite extension E/B that is a splitting field of some polynomial $f(x) \in k[x]$ (such an extension of smallest degree is called a **normal**[4] **closure** of B/k). Moreover, if each α_i is separable over k, then $f(x)$ can be chosen to be a separable polynomial.*

(ii) *If B is a radical extension of k, then the extension E/B in part (i) is a radical extension of k.*

Proof. (i) By Theorem 3.120(i), there is an irreducible polynomial $p_i(x) = \mathrm{irr}(\alpha_i, k)$ in $k[x]$, for each i, with $p_i(\alpha_i) = 0$, and a splitting field E of $f(x) = p_1(x)\cdots p_n(x)$ containing B. If each α_i is separable over k, then each $p_i(x)$ is a separable polynomial, and hence $f(x)$ is a separable polynomial.

(ii) For each pair of roots α and α' of any $p_i(x)$, there is an isomorphism $\gamma\colon k(\alpha) \to k(\alpha')$ which fixes k and which takes $\alpha \mapsto \alpha'$, for both $k(\alpha)$ and $k(\alpha')$ are isomorphic to $k[x]/(p_i(x))$. By Lemma 3.130, each such γ extends to an automorphism $\sigma \in G = \mathrm{Gal}(E/k)$. It follows that $E = k(\sigma(u_1), \ldots, \sigma(u_t) : \sigma \in G)$.

If B/k is a radical extension, then

$$k \subseteq k(u_1) \subseteq k(u_1, u_2) \subseteq \cdots \subseteq k(u_1, \ldots, u_t) = B,$$

where each $k(u_1, \ldots, u_{i+1})$ is a pure extension of $k(u_1, \ldots, u_i)$; of course, $\sigma(B) = k(\sigma(u_1), \ldots, \sigma(u_t))$ is a radical extension of k for every $\sigma \in G$. We now show that E is a radical extension of k. Define

$$B_1 = k(\sigma(u_1) : \sigma \in G).$$

Now if $G = \{1, \sigma, \tau, \ldots\}$, then the tower

$$k \subseteq k(u_1) \subseteq k(u_1, \sigma(u_1)) \subseteq k(u_1, \sigma(u_1), \tau(u_1)) \subseteq \cdots \subseteq B_1$$

displays B_1 as a radical extension of k. For example, if u_1^m lies in k, then $\tau(u_1)^m = \tau(u_1^m)$ lies in $\tau(k) = k$, and hence $\tau(u_1)^m$ lies in $k \subseteq k(u_1, \sigma(u_1))$. Assuming, by induction, that

[4]We often call an extension E/k a ***normal extension*** if it is the splitting field of some set of polynomials in $k[x]$.

a radical extension B_i/k containing $\{\sigma(u_j) : \sigma \in G\}$ for all $j \leq i$ has been constructed, define

$$B_{i+1} = B_i(\sigma(u_{i+1}) : \sigma \in G).$$

It is easy to see that B_{i+1}/B_i is a radical extension: If $u_{i+1}^m \in k(u_1, \ldots, u_i)$, then $\tau(u_{i+1})^m \in k(\tau(u_1), \ldots, \tau(u_i)) \subseteq B_i$; it follows that B_{i+1} is a radical extension of k. Finally, since $E = B_t$, we have shown that E is a radical extension of k. •

We can now give the heart of the translation we have been seeking.

Lemma 4.18. *Let*

$$K_0 \subseteq K_1 \subseteq K_2 \subseteq \cdots \subseteq K_t$$

be a radical extension of a field K_0. Assume, for each $i \geq 1$, that each K_i is a pure extension of prime type p_i over K_{i-1}, where $p_i \neq \mathrm{char}(K_0)$, and that K_0 contains all the p_ith roots of unity. If K_t is a splitting field over K_0, then there is a sequence of subgroups

$$\mathrm{Gal}(K_t/K_0) = G_0 \geq G_1 \geq G_2 \geq \cdots \geq G_t = \{1\},$$

with each G_{i+1} a normal subgroup of G_i and with G_i/G_{i+1} cyclic of prime order p_{i+1}.

Proof. For each i, define $G_i = \mathrm{Gal}(K_t/K_i)$. It is clear that

$$\mathrm{Gal}(K_t/K_0) = G_0 \geq G_1 \geq G_2 \geq \cdots \geq G_t = \{1\}$$

is a sequence of subgroups. Since $K_1 = K_0(u)$, where $u^{p_1} \in K_0$, the assumptions that $\mathrm{char}(K_0) \neq p_1$ and that K_0 contains all the p_1th roots of unity implies that K_0 contains a primitive p_1th root of unity ω; hence, K_1 is a splitting field of the separable polynomial $x^{p_1} - u^{p_1}$, for the roots are $u, \omega u, \ldots, \omega^{p_1-1}u$. We may thus apply Theorem 4.16 to see that $G_1 = \mathrm{Gal}(K_t/K_1)$ is a normal subgroup of $G_0 = \mathrm{Gal}(K_t/K_0)$ and that $G_0/G_1 \cong \mathrm{Gal}(K_1/K_0)$. By Theorem 4.7(ii), $G_0/G_1 \cong \mathbb{I}_{p_1}$. This argument can be repeated for each i. •

We have been led to the following definition.

Definition. A ***normal series***[5] of a group G is a sequence of subgroups

$$G = G_0 \geq G_1 \geq G_2 \geq \cdots \geq G_t = \{1\}$$

with each G_{i+1} a normal subgroup of G_i; the ***factor groups*** of this series are the quotient groups

$$G_0/G_1, G_1/G_2, \ldots, G_{n-1}/G_n.$$

A finite group G is called ***solvable*** if it has a normal series each of whose factor groups has prime order (see the definition of infinite solvable groups on page 286).

[5]This terminology is not quite standard. We know that normality is not transitive; that is, if $H \leq K$ are subgroups of a group G, then $H \lhd K$ and $K \lhd G$ does not force $H \lhd G$. A subgroup $H \leq G$ is called a ***subnormal subgroup*** if there is a chain

$$G = G_0 \geq G_1 \geq G_2 \geq \cdots \geq G_t = H$$

with $G_i \lhd G_{i-1}$ for all $i \geq 1$. Normal series as defined in the text are called ***subnormal series*** by some authors; they reserve the name *normal series* for those series in which each G_i is a normal subgroup of the big group G.

In this language, Lemma 4.18 says that $\mathrm{Gal}(K_t/K_0)$ is a solvable group if K_t is a radical extension of K_0 and K_0 contains appropriate roots of unity.

Example 4.19.
(i) By Exercise 2.86(ii) on page 113, every finite abelian group G has a (necessarily normal) subgroup of prime index. It follows, by induction on $|G|$, that every finite abelian group is solvable.

(ii) Let us see that S_4 is a solvable group. Consider the chain of subgroups

$$S_4 \geq A_4 \geq \mathbf{V} \geq W \geq \{1\},$$

where $\mathbf{V}$ is the four-group and W is any subgroup of $\mathbf{V}$ of order 2. Note, since $\mathbf{V}$ is abelian, that W is a normal subgroup of $\mathbf{V}$. Now $|S_4/A_4| = |S_4|/|A_4| = 24/12 = 2$, $|A_4/\mathbf{V}| = |A_4|/|\mathbf{V}| = 12/4 = 3$, $|\mathbf{V}/W| = |\mathbf{V}|/|W| = 4/2 = 2$, and $|W/\{1\}| = |W| = 2$. Since each factor group has prime order, S_4 is solvable.

(iii) A nonabelian simple group G, for example, $G = A_5$, is not solvable, for its only proper normal subgroup is $\{1\}$, and $G/\{1\} \cong G$ is not cyclic of prime order. ◀

The awkward hypothesis in the next lemma, about roots of unity, will soon be removed.

Lemma 4.20. *Let k be a field and let $f(x) \in k[x]$ be solvable by radicals, so there is a radical extension $k = K_0 \subseteq K_1 \subseteq \cdots \subseteq K_t$ with K_t containing a splitting field E of $f(x)$. If each K_i/K_{i-1} is a pure extension of prime type p_i, where $p_i \neq char(k)$, and if k contains all the p_ith roots of unity, then the Galois group* $\mathrm{Gal}(E/k)$ *is a quotient of a solvable group.*

Proof. There is a tower of pure extensions of prime type

$$k = K_0 \subseteq K_1 \subseteq K_2 \subseteq \cdots \subseteq K_t$$

with $E \subseteq K_t$; by Lemma 4.17, we may assume that K_t is also a splitting field of some polynomial in $k[x]$. The hypothesis on k allows us to apply Lemma 4.18 to see that $\mathrm{Gal}(K_t/k)$ is a solvable group. Since E and K_t are splitting fields over k, Theorem 4.16 shows that $\mathrm{Gal}(K_t/k)/\mathrm{Gal}(K_t/E) \cong \mathrm{Gal}(E/k)$, as desired. •

Proposition 4.21. *Every quotient G/N of a solvable group G is itself a solvable group.*

Proof. Let $G = G_0 \geq G_1 \geq G_2 \geq \cdots \geq G_t = \{1\}$ be a sequence of subgroups as in the definition of solvable group. Since $N \lhd G$, we have NG_i a subgroup of G for all i, and so there is a sequence of subgroups

$$G = G_0N \geq G_1N \geq \cdots \geq G_tN = N \geq \{1\}.$$

This is a normal series: With obvious notation,

$$(g_in)G_{i+1}N(g_in)^{-1} \leq g_iG_{i+1}Ng_i^{-1} = g_iG_{i+1}g_i^{-1}N \leq G_{i+1}N;$$

the first inequality holds because $n(G_{i+1}N)n^{-1} \leq NG_{i+1}N \leq (G_{i+1}N)(G_{i+1}N) = G_{i+1}N$ (for $G_{i+1}N$ is a subgroup); the equality holds because $Ng_i^{-1} = g_i^{-1}N$ (for $N \lhd G$, and so its right cosets coincide with its left cosets); the last inequality holds because $G_{i+1} \lhd G_i$.

The second isomorphism theorem gives

$$\frac{G_i}{G_i \cap (G_{i+1}N)} \cong \frac{G_i(G_{i+1}N)}{G_{i+1}N} = \frac{G_i N}{G_{i+1}N},$$

the last equation holding because $G_i G_{i+1} = G_i$. Since $G_{i+1} \lhd G_i \cap G_{i+1}N$, the third isomorphism theorem gives a surjection $G_i/G_{i+1} \to G_i/[G_i \cap G_{i+1}N]$, and so the composite is a surjection $G_i/G_{i+1} \to G_iN/G_{i+1}N$. As G_i/G_{i+1} is cyclic of prime order, its image is either cyclic of prime order or trivial. Therefore, G/N is a solvable group. •

Proposition 4.22. *Every subgroup H of a solvable group G is itself a solvable group.*

Proof. Since G is solvable, there is a sequence of subgroups

$$G = G_0 \geq G_1 \geq G_2 \geq \cdots \geq G_t = \{1\}$$

with G_i normal in G_{i-1} and G_{i-1}/G_i cyclic, for all i. Consider the sequence of subgroups

$$H = H \cap G_0 \geq H \cap G_1 \geq H \cap G_2 \geq \cdots \geq H \cap G_t = \{1\}.$$

This is a normal series: If $h_{i+1} \in H \cap G_{i+1}$ and $g_i \in H \cap G_i$, then $g_i h_{i+1} g_i^{-1} \in H$, for $g_i, h_{i+1} \in H$; also, $g_i h_{i+1} g_i^{-1} \in G_{i+1}$ because G_{i+1} is normal in G_i. Therefore, $g_i h_{i+1} g_i^{-1} \in H \cap G_{i+1}$, and so $H \cap G_{i+1} \lhd H \cap G_i$. Finally, the second isomorphism theorem gives

$$\begin{aligned}(H \cap G_i)/(H \cap G_{i+1}) &= (H \cap G_i)/[(H \cap G_i) \cap G_{i+1}] \\ &\cong G_{i+1}(H \cap G_i)/G_{i+1}.\end{aligned}$$

But the last (quotient) group is a subgroup of G_i/G_{i+1}. Since the only subgroups of a cyclic group C of prime order are C and $\{1\}$, it follows that the nontrivial factor groups $(H \cap G_i)/(H \cap G_{i+1})$ are cyclic of prime order. Therefore, H is a solvable group. •

Example 4.23.
In Example 4.19(ii), we showed that S_4 is a solvable group. However, if $n \geq 5$, the symmetric group S_n is not a solvable group. If, on the contrary, S_n were solvable, then so would each of its subgroups be solvable. But $A_5 \leq S_5 \leq S_n$, and A_5 is not solvable because it is a nonabelian simple group. ◂

Proposition 4.24. *If $H \lhd G$ and if both H and G/H are solvable groups, then G is solvable.*

Proof. Since G/H is solvable, there is a normal series

$$G/H \geq K_1^* \geq K_2^* \geq \cdots K_m^* = \{1\}$$

having factor groups of prime order. By the correspondence theorem for groups, there are subgroups K_i of G,

$$G \geq K_1 \geq K_2 \geq \cdots \geq K_m = H,$$

with $K_i/H = K_i^*$ and $K_{i+1} \lhd K_i$ for all i. By the third isomorphism theorem,

$$K_i^*/K_{i+1}^* \cong K_i/K_{i+1}$$

for all i, and so K_i/K_{i+1} is cyclic of prime order for all i.

Since H is solvable, there is a normal series

$$H \geq H_1 \geq H_2 \geq \cdots H_q = \{1\}$$

having factor groups of prime order. Splice these two series together,

$$G \geq K_1 \geq K_2 \geq \cdots \geq K_m \geq H_1 \geq H_2 \geq \cdots H_q = \{1\},$$

to obtain a normal series of G having factor groups of prime order. •

Corollary 4.25. *If H and K are solvable groups, then $H \times K$ is solvable.*

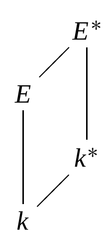
Proof. Since $(H \times K)/H \cong K$, the result follows at once from Proposition 4.24. •

We return to fields, for we can now give the main criterion that a polynomial be solvable by radicals.

Theorem 4.26 (*Galois*). *Let $f(x) \in k[x]$, where k is a field, and let E be a splitting field of $f(x)$ over k. If $f(x)$ is solvable by radicals, then its Galois group* $\text{Gal}(E/k)$ *is a solvable group.*

Remark. The converse of this theorem is false if k has characteristic $p > 0$ (see Proposition 4.56), but it is true when k has characteristic 0 (see Theorem 4.53). ◂

Proof. In the proof of Lemma 4.20, we assumed that the ground field contained certain p_ith roots of unity (the primes p_i were types of pure extensions). Define m to be the product of all these p_i, define E^* to be a splitting field of $x^m - 1$ over E, and define $k^* = k(\Omega)$, where Ω is the set of all mth roots of unity in E^*. Now E^* is a splitting field of $f(x)$ over k^*, and so $\text{Gal}(E^*/k^*)$ is solvable, by Proposition 4.21.



Consider the tower $k \subseteq k^* \subseteq E^*$; we have $\mathrm{Gal}(E^*/k^*) \lhd \mathrm{Gal}(E^*/k)$, by Theorem 4.16, and

$$\mathrm{Gal}(E^*/k)/\,\mathrm{Gal}(E^*/k^*) \cong \mathrm{Gal}(k^*/k).$$

Now $\mathrm{Gal}(E^*/k^*)$ is solvable, while $\mathrm{Gal}(k^*/k)$ is abelian, hence solvable, by Proposition 4.11; therefore, $\mathrm{Gal}(E^*/k)$ is solvable, by Proposition 4.24. Finally, we may use Theorem 4.16 once again, for the tower $k \subseteq E \subseteq E^*$ satisfies the hypothesis that both E and E^* are splitting fields of polynomials in $k[x]$ [E^* is a splitting field of $(x^m - 1)f(x)$]. It follows that $\mathrm{Gal}(E^*/k)/\,\mathrm{Gal}(E^*/E) \cong \mathrm{Gal}(E/k)$, and so $\mathrm{Gal}(E/k)$ is solvable, for it is a quotient of a solvable group. •

Recall that if k is a field and $E = k(y_1, \ldots, y_n) = \mathrm{Frac}(k[y_1, \ldots, y_n])$ is the field of rational functions, then the *general polynomial of degree n* over k is

$$(x - y_1)(x - y_2)\cdots(x - y_n).$$

Galois's theorem is strong enough to prove that there is no generalization of the quadratic formula for the general quintic polynomial.

Theorem 4.27 (***Abel–Ruffini***). *If $n \geq 5$, the general polynomial of degree n*

$$f(x) = (x - y_1)(x - y_2)\cdots(x - y_n)$$

over a field k is not solvable by radicals.

Proof. In Example 3.125, we saw that if $E = k(y_1, \ldots, y_n)$ is the field of all rational functions in n variables with coefficients in a field k, and if $F = k(a_0, \ldots, a_{n-1})$, where the a_i are the coefficients of $f(x)$, then E is the splitting field of $f(x)$ over F.

We claim that $\mathrm{Gal}(E/F) \cong S_n$. Exercise 3.47(i) on page 150 says that if A and R are domains and $\varphi\colon A \to R$ is an isomorphism, then $a/b \mapsto \varphi(a)/\varphi(b)$ is an isomorphism $\mathrm{Frac}(A) \to \mathrm{Frac}(R)$. In particular, if $\sigma \in S_n$, then there is an automorphism $\widetilde{\sigma}$ of $k[y_1, \ldots, y_n]$ defined by $\widetilde{\sigma}\colon f(y_1, \ldots, y_n) \mapsto f(y_{\sigma 1}, \ldots, y_{\sigma n})$; that is, $\widetilde{\sigma}$ just permutes the variables, and $\widetilde{\sigma}$ extends to an automorphism σ^* of $E = \mathrm{Frac}(k[y_1, \ldots, y_n])$. Equations (1) on page 198 show that σ^* fixes F, and so $\sigma^* \in \mathrm{Gal}(E/F)$. Using Lemma 4.2, it is easy to see that $\sigma \mapsto \sigma^*$ is an injection $S_n \to \mathrm{Gal}(E/F)$, so that $|S_n| \leq |\,\mathrm{Gal}(E/F)|$. On the other hand, Theorem 4.3 shows that $\mathrm{Gal}(E/F)$ can be imbedded in S_n, giving the reverse inequality $|\,\mathrm{Gal}(E/F)| \leq |S_n|$. Therefore, $\mathrm{Gal}(E/F) \cong S_n$. But S_n is not a solvable group if $n \geq 5$, by Example 4.23, and so Theorem 4.26 shows that $f(x)$ is not solvable by radicals. •

We know that some quintics in $\mathbb{Q}[x]$ are solvable by radicals; for example, $x^5 - 1$ is solvable by radicals, for its Galois group is abelian, by Proposition 4.11. On the other hand, we can give specific quintics in $\mathbb{Q}[x]$ that are not solvable by radicals. For example, $f(x) = x^5 - 4x + 2 \in \mathbb{Q}[x]$ is not solvable by radicals, for it can be shown that its Galois group is isomorphic to S_5 (see Exercise 4.13 on page 218).

EXERCISES

4.1 Given $u, v \in \mathbb{C}$, prove that there exist $g, h \in \mathbb{C}$ with $u = g + h$ and $v = gh$.

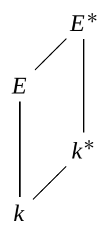
4.2 Show that the quadratic formula does not hold for $ax^2 + bx + c \in k[x]$ when characteristic(k) = 2.

4.3 (i) Find the roots of $f(x) = x^3 - 3x + 1 \in \mathbb{Q}[x]$.

(ii) Find the roots of $f(x) = x^4 - 2x^2 + 8x - 3 \in \mathbb{Q}[x]$.

4.4 Let $f(x) \in E[x]$, where E is a field, and let $\sigma : E \to E$ be an automorphism. If $f(x)$ splits and σ fixes every root of $f(x)$, prove that σ fixes every coefficient of $f(x)$.

4.5 (***Accessory Irrationalities***) Let E/k be a splitting field of $f(x) \in k[x]$ with Galois group $G = \text{Gal}(E/k)$. Prove that if k^*/k is a field extension and E^* is a splitting field



of $f(x)$ over k^*, then restriction, $\sigma \mapsto \sigma|E$, is an injective homomorphism

$$\text{Gal}(E^*/k^*) \to \text{Gal}(E/k).$$

Hint. If $\sigma \in \text{Gal}(E^*/k^*)$, then σ permutes the roots of $f(x)$, so that $\sigma|E \in \text{Gal}(E/k)$.

4.6 (i) Let K/k be a field extension, and let $f(x) \in k[x]$ be a separable polynomial. Prove that $f(x)$ is a separable polynomial when viewed as a polynomial in $K[x]$.

(ii) Let k be a field, and let $f(x), g(x) \in k[x]$. Prove that if both $f(x)$ and $g(x)$ are separable polynomials, then their product $f(x)g(x)$ is also a separable polynomial.

4.7 Let k be a field and let $f(x) \in k[x]$ be a separable polynomial. If E/k is a splitting field of $f(x)$, prove that every root of $f(x)$ in E is a separable element over k.

4.8 Let K/k be a field extension that is a splitting field of a polynomial $f(x) \in k[x]$. If $p(x) \in k[x]$ is a monic irreducible polynomial with no repeated roots, and if

$$p(x) = g_1(x) \cdots g_r(x) \quad \text{in} \quad K[x],$$

where the $g_i(x)$ are monic irreducible polynomials in $K[x]$, prove that all the $g_i(x)$ have the same degree. Conclude that $\deg(p) = r \deg(g_i)$.

Hint. In some splitting field E/K of $p(x)f(x)$, let α be a root of $g_i(x)$ and β be a root of $g_j(x)$, where $i \neq j$. There is an isomorphism $\varphi \colon k(\alpha) \to k(\beta)$ with $\varphi(\alpha) = \beta$, which fixes k and which admits an extension to $\Phi \colon E \to E$. Show that $\Phi|K$ induces an automorphism of $K[x]$ taking $g_i(x)$ to $g_j(x)$.

4.9 (i) Give an example of a group G having a subnormal subgroup that is not a normal subgroup.

(ii) Give an example of a group G having a subgroup that is not a subnormal subgroup.

4.10 Prove that the following statements are equivalent for a quadratic $f(x) = ax^2 + bx + c \in \mathbb{Q}[x]$.

(i) $f(x)$ is irreducible in $\mathbb{Q}[x]$.

(ii) $\sqrt{b^2 - 4ac}$ is not rational.

(iii) $\mathrm{Gal}(\mathbb{Q}(\sqrt{b^2 - 4ac}), \mathbb{Q})$ has order 2.

4.11 Let k be a field, let $f(x) \in k[x]$ be a polynomial of degree p, where p is prime, and let E/k be a splitting field. Prove that if $\mathrm{Gal}(E/k) \cong \mathbb{I}_p$, then $f(x)$ is irreducible.
Hint. Show that $f(x)$ has no repeated roots.

4.12 (i) Prove that if σ is a 5-cycle and τ is a transposition, then S_5 is generated by $\{\sigma, \tau\}$.
Hint. Use Exercise 2.94(iii) on page 114.

(ii) Give an example showing that S_n, for some n, contains an n-cycle σ and a transposition τ such that $\langle \sigma, \tau \rangle \neq S_n$.

4.13 Let $f(x) = x^5 - 4x + 2 \in \mathbb{Q}[x]$ and let G be its Galois group.

(i) Assuming that $f(x)$ is an irreducible polynomial, prove that $|G|$ is a multiple of 5. [We can prove that $f(x)$ is irreducible using Eisenstein's criterion, Theorem 6.34 on page 337.]

(ii) Prove that $f(x)$ has three real roots and two complex roots, which are, of course, complex conjugates. Conclude that if the Galois group G of $f(x)$ is viewed as a subgroup of S_5, then G contains complex conjugation, which is a transposition of the roots of $f(x)$.

(iii) Prove that $G \cong S_5$, and conclude that $f(x)$ is not solvable by radicals.
Hint. Use Exercise 4.12.

4.2 FUNDAMENTAL THEOREM OF GALOIS THEORY

Galois theory analyzes the connection between algebraic extensions E of a field k and the corresponding Galois groups $\mathrm{Gal}(E/k)$. This connection will enable us to prove the converse of Galois's theorem: If k is a field of characteristic 0, and if $f(x) \in k[x]$ has a solvable Galois group, then $f(x)$ is solvable by radicals. The fundamental theorem of algebra is also a consequence of this analysis.

We have already seen several theorems about Galois groups whose hypothesis involves an extension being a splitting field of some polynomial. Let us begin by asking whether there is some intrinsic property of an extension E/k that characterizes its being a splitting field, without referring to any particular polynomial in $k[x]$. It turns out that the way to understand splitting fields E/k is to examine them in the context of both separability and the action of the Galois group $\mathrm{Gal}(E/k)$ on E.

Let E be a field and let $\mathrm{Aut}(E)$ be the group of all (field) automorphisms of E. If k is any subfield of E, then $\mathrm{Gal}(E/k)$ is a subgroup of $\mathrm{Aut}(E)$, and so it acts on E. Whenever a group acts on a set, we are interested in its orbits and stabilizers, but we now ask for those elements of E stabilized by every σ in some subset H of $\mathrm{Aut}(E)$.

Definition. If E is a field and H is a subset of $\mathrm{Aut}(E)$, then the ***fixed field*** of H is defined by
$$E^H = \{a \in E : \sigma(a) = a \text{ for all } \sigma \in H\}.$$

The most important instance of a fixed field E^H arises when H is a subgroup of $\mathrm{Aut}(E)$, but we will meet a case in which it is merely a subset.

It is easy to see that if $\sigma \in \mathrm{Aut}(E)$, then $E^\sigma = \{a \in E : \sigma(a) = a\}$ is a subfield of E; it follows that E^H is a subfield of E, for

$$E^H = \bigcap_{\sigma \in H} E^\sigma.$$

In Example 3.125, we considered $E = k(y_1, \ldots, y_n)$, the rational function field in n variables with coefficients in a field k, and its subfield $K = k(a_0, \ldots, a_{n-1})$, where

$$f(x) = (x - y_1)(x - y_2) \cdots (x - y_n) = a_0 + a_1 x + \cdots + a_{n-1}x^{n-1} + x^n$$

is the general polynomial of degree n over k. We saw that E is a splitting field of $f(x)$ over K, for it arises from K by adjoining to it all the roots of $f(x)$, namely, all the y's. Now the symmetric group $S_n \le \mathrm{Aut}(E)$, for every permutation of $y_1, \ldots, y_n$ extends to an automorphism of E, and it turns out that $K = E^{S_n}$. The elements of K are usually called the *symmetric functions* in n variables over k.

Definition. A rational function $g(x_1, \ldots, x_n)/h(x_1, \ldots, x_n) \in k(x_1, \ldots, x_n)$ is a ***symmetric function*** if it is unchanged by permuting its variables: For every $\sigma \in S_n$, we have $g(x_{\sigma 1}, \ldots, x_{\sigma n})/h(x_{\sigma 1}, \ldots, x_{\sigma n}) = g(x_1, \ldots, x_n)/h(x_1, \ldots, x_n)$.

The various polynomials in Eqs. (1) on page 198 define examples of symmetric functions; they are called the ***elementary symmetric functions***.

The proof of the following proposition is almost obvious.

Proposition 4.28. *If E is a field, then the function $H \mapsto E^H$, from subsets H of* $\mathrm{Aut}(E)$ *to subfields of E, is* ***order-reversing***: *If $H \le L \le \mathrm{Aut}(E)$, then $E^L \subseteq E^H$.*

Proof. If $a \in E^L$, then $\sigma(a) = a$ for all $\sigma \in L$. Since $H \le L$, it follows, in particular, that $\sigma(a) = a$ for all $\sigma \in H$. Hence, $E^L \subseteq E^H$. •

Example 4.29.
Suppose now that k is a subfield of E and that $G = \mathrm{Gal}(E/k)$. It is obvious that $k \subseteq E^G$, but the inclusion can be strict. For example, let $E = \mathbb{Q}(\sqrt[3]{2})$. If $\sigma \in G = \mathrm{Gal}(E/\mathbb{Q})$, then σ must fix $\mathbb{Q}$, and so it permutes the roots of $f(x) = x^3 - 2$. But the other two roots of $f(x)$ are not real, so that $\sigma(\sqrt[3]{2}) = \sqrt[3]{2}$. It now follows from Lemma 4.2 that σ is the identity; that is, $E^G = E$. Note that E is not a splitting field of $f(x)$. ◂

Our immediate goal is to determine the degree $[E : E^G]$, where $G \le \mathrm{Aut}(E)$. To this end, we introduce the notion of characters.

Definition. A ***character***[6] of a group G in a field E is a (group) homomorphism $\sigma\colon G \to E^\times$, where $E^\times$ denotes the multiplicative group of nonzero elements of the field E.

If $\sigma \in \operatorname{Aut}(E)$, then its restriction $\sigma|E^\times\colon E^\times \to E^\times$ is a character in E.

Definition. If E is a field and $G \le \operatorname{Aut}(E)$, then a list $\sigma_1, \ldots, \sigma_n$ of characters of G in E is ***independent*** if, whenever $c_1, \ldots, c_n \in E$ and

$$\sum_i c_i\sigma_i(x) = 0 \qquad \text{for all } x \in G,$$

then all the $c_i = 0$.

In Example 3.82(iii), we saw that the set E^X of all the functions from a set X to a field E is a vector space over E, where addition of functions is defined by

$$\sigma + \tau\colon x \mapsto \sigma(x) + \tau(x),$$

and scalar multiplication is defined, for $c \in E$, by

$$c\sigma\colon x \mapsto c\sigma(x).$$

Independence of characters, as just defined, is linear independence in the vector space E^X when X is the group G.

Proposition 4.30 (Dedekind). *Every list $\sigma_1, \ldots, \sigma_n$ of distinct characters of a group G in a field E is independent.*

Proof. The proof is by induction on $n \ge 1$. The base step $n = 1$ is true, for if $c\sigma(x) = 0$ for all $x \in G$, then either $c = 0$ or $\sigma(x) = 0$; but $\sigma(x) \ne 0$, because $\operatorname{im}\sigma \subseteq E^\times$.

Assume that $n > 1$; if the characters are not independent, there are $c_i \in E$, not all zero, with

$$c_1\sigma_1(x) + \cdots + c_{n-1}\sigma_{n-1}(x) + c_n\sigma_n(x) = 0 \tag{2}$$

for all $x \in G$. We may assume that all $c_i \ne 0$, or we may invoke the inductive hypothesis and reach a contradiction, as desired. Multiplying by c_n^{-1} if necessary, we may assume that $c_n = 1$. Since $\sigma_n \ne \sigma_1$, there exists $y \in G$ with $\sigma_1(y) \ne \sigma_n(y)$. In Eq. (2), replace x by yx to obtain

$$c_1\sigma_1(y)\sigma_1(x) + \cdots + c_{n-1}\sigma_{n-1}(y)\sigma_{n-1}(x) + \sigma_n(y)\sigma_n(x) = 0,$$

[6] This definition is a special case of *character* in representation theory: If $\sigma\colon G \to \mathrm{GL}(n, E)$ is a homomorphism, then its ***character*** $\chi_\sigma\colon G \to E$ is defined, for $x \in G$, by

$$\chi_\sigma(x) = \operatorname{trace}(\sigma(x)),$$

where the *trace* of an $n \times n$ matrix is the sum of its diagonal entries. When $n = 1$, then $\mathrm{GL}(1, E) = E^\times$ and $\chi_\sigma(x) = \sigma(x)$ is called a ***linear character***.

for $\sigma_i(yx) = \sigma_i(y)\sigma_i(x)$. Now multiply this equation by $\sigma_n(y)^{-1}$ to obtain the equation

$$c_1\sigma_n(y)^{-1}\sigma_1(y)\sigma_1(x) + \cdots + c_{n-1}\sigma_n(y)^{-1}\sigma_{n-1}(y)\sigma_{n-1}(x) + \sigma_n(x) = 0.$$

Subtract this last equation from Eq. (2) to obtain a sum of $n - 1$ terms:

$$c_1[1 - \sigma_n(y)^{-1}\sigma_1(y)]\sigma_1(x) + c_2[1 - \sigma_n(y)^{-1}\sigma_2(y)]\sigma_2(x) + \cdots = 0.$$

By induction, each of the coefficients $c_i[1 - \sigma_n(y)^{-1}\sigma_i(y)] = 0$. Now $c_i \neq 0$, and so $\sigma_n(y)^{-1}\sigma_i(y) = 1$ for all $i < n$. In particular, $\sigma_n(y) = \sigma_1(y)$, contradicting the definition of y. •

Lemma 4.31. *If $G = \{\sigma_1, \ldots, \sigma_n\}$ is a set of n distinct automorphisms of a field E, then*

$$[E : E^G] \geq n.$$

Proof. Suppose, on the contrary, that $[E : E^G] = r < n$, and let $\alpha_1, \ldots, \alpha_r$ be a basis of E/E^G. Consider the homogeneous linear system over E of r equations in n unknowns:

$$\begin{aligned}
\sigma_1(\alpha_1)x_1 + \cdots + \sigma_n(\alpha_1)x_n &= 0\\
\sigma_1(\alpha_2)x_1 + \cdots + \sigma_n(\alpha_2)x_n &= 0\\
\vdots \qquad \vdots \qquad\qquad &\vdots\\
\sigma_1(\alpha_r)x_1 + \cdots + \sigma_n(\alpha_r)x_n &= 0.
\end{aligned}$$

Since $r < n$, there are fewer equations than variables, and so there is a nontrivial solution $(c_1, \ldots, c_n)$ in E^n.

We are now going to show that $\sigma_1(\beta)c_1 + \cdots + \sigma_n(\beta)c_n = 0$ for any $\beta \in E^\times$, which will contradict the independence of the characters $\sigma_1|E^\times, \ldots, \sigma_n|E^\times$. Since $\alpha_1, \ldots, \alpha_r$ is a basis of E over E^G, every $\beta \in E$ can be written

$$\beta = \sum b_i\alpha_i,$$

where $b_i \in E^G$. Multiply the ith row of the system by $\sigma_1(b_i)$ to obtain the system with ith row:

$$\sigma_1(b_i)\sigma_1(\alpha_i)c_1 + \cdots + \sigma_1(b_i)\sigma_n(\alpha_i)c_n = 0.$$

But $\sigma_1(b_i) = b_i = \sigma_j(b_i)$ for all i, j, because $b_i \in E^G$. Thus, the system has ith row:

$$\sigma_1(b_i\alpha_i)c_1 + \cdots + \sigma_n(b_i\alpha_i)c_n = 0.$$

Adding all the rows gives

$$\sigma_1(\beta)c_1 + \cdots + \sigma_n(\beta)c_n = 0,$$

which contradicts the independence of the characters $\sigma_1, \ldots, \sigma_n$. •

Proposition 4.32. *If $G = \{\sigma_1, \dots, \sigma_n\}$ is a subgroup of* Aut(E), *then*

$$[E : E^G] = |G|.$$

Proof. In light of Lemma 4.31, it suffices to prove $[E : E^G] \le |G|$. If, on the contrary, $[E : E^G] > n$, let $\{\omega_1, \dots, \omega_{n+1}\}$ be a linearly independent list of vectors in E over E^G. Consider the system of n equations in $n+1$ unknowns:

$$\begin{aligned} \sigma_1(\omega_1)x_1 + \cdots + \sigma_1(\omega_{n+1})x_{n+1} &= 0 \\ \vdots \qquad\qquad\qquad & \vdots \\ \sigma_n(\omega_1)x_1 + \cdots + \sigma_n(\omega_{n+1})x_{n+1} &= 0. \end{aligned}$$

There is a nontrivial solution $(\alpha_1, \dots, \alpha_{n+1})$ over E; we proceed to normalize it. Choose a solution $(\beta_1, \dots, \beta_r, 0, \dots, 0)$ having the smallest number r of nonzero components (by reindexing the ω_i, we may assume that all nonzero components come first). Note that $r \ne 1$, lest $\sigma_1(\omega_1)\beta_1 = 0$ imply $\beta_1 = 0$. Multiplying by its inverse if necessary, we may assume that $\beta_r = 1$. Not all $\beta_i \in E^G$, lest the row corresponding to $\sigma = 1_E$ violates the linear independence of $\{\omega_1, \dots, \omega_{n+1}\}$. Our last assumption is that β_1 does not lie in E^G (this, too, can be accomplished by reindexing the ω_i). There thus exists σ_k with $\sigma_k(\beta_1) \ne \beta_1$. Since $\beta_r = 1$, the original system has jth row

$$\sigma_j(\omega_1)\beta_1 + \cdots + \sigma_j(\omega_{r-1})\beta_{r-1} + \sigma_j(\omega_r) = 0. \tag{3}$$

Apply σ_k to this system to obtain

$$\sigma_k\sigma_j(\omega_1)\sigma_k(\beta_1) + \cdots + \sigma_k\sigma_j(\omega_{r-1})\sigma_k(\beta_{r-1}) + \sigma_k\sigma_j(\omega_r) = 0.$$

Since G is a group, $\sigma_k\sigma_1, \dots, \sigma_k\sigma_n$ is just a permutation of $\sigma_1, \dots, \sigma_n$. Setting $\sigma_k\sigma_j = \sigma_i$, the system has ith row

$$\sigma_i(\omega_1)\sigma_k(\beta_1) + \cdots + \sigma_i(\omega_{r-1})\sigma_k(\beta_{r-1}) + \sigma_i(\omega_r) = 0.$$

Subtract this from the ith row of Eq. (3) to obtain a new system with ith row:

$$\sigma_i(\omega_1)[\beta_1 - \sigma_k(\beta_1)] + \cdots + \sigma_i(\omega_{r-1})[\beta_{r-1} - \sigma_k(\beta_{r-1})] = 0.$$

Since $\beta_1 - \sigma_k(\beta_1) \ne 0$, we have found a nontrivial solution of the original system having fewer than r nonzero components, a contradiction. •

These ideas give a result needed in the proof of the fundamental theorem of Galois theory.

Theorem 4.33. *If G and H are finite subgroups of* Aut(E) *with $E^G = E^H$, then $G = H$.*

Proof. We first show that if $\sigma \in \text{Aut}(E)$, then σ fixes E^G if and only if $\sigma \in G$. Clearly, σ fixes E^G if $\sigma \in G$. Suppose, conversely, that σ fixes E^G but $\sigma \notin G$. If $|G| = n$, then

$$n = |G| = [E : E^G],$$

by Proposition 4.32. Since σ fixes E^G, we have $E^G \subseteq E^{G \cup \{\sigma\}}$. But the reverse inequality always holds, by Proposition 4.28, so that $E^G = E^{G \cup \{\sigma\}}$. Hence,

$$n = [E : E^G] = [E : E^{G \cup \{\sigma\}}] \geq |G \cup \{\sigma\}| = n + 1,$$

by Lemma 4.31, giving the contradiction $n \geq n + 1$.

If $\sigma \in H$, then σ fixes $E^H = E^G$, and hence $\sigma \in G$; that is, $H \leq G$; the reverse inclusion is proved the same way, and so $H = G$. •

We can now give the characterization of splitting fields we have been seeking.

Theorem 4.34. *If E/k is a finite extension with Galois group $G = \text{Gal}(E/k)$, then the following statements are equivalent.*

(i) *E is a splitting field of some separable polynomial $f(x) \in k[x]$.*

(ii) $k = E^G$.

(iii) *Every irreducible $p(x) \in k[x]$ having one root in E is separable and splits in $E[x]$.*

Proof. (i) ⇒ (ii) By Theorem 4.7(ii), $|G| = [E : k]$. But Proposition 4.32 gives $|G| = [E : E^G]$, so that

$$[E : k] = [E : E^G].$$

Since $k \leq E^G$, we have $[E : k] = [E : E^G][E^G : k]$, so that $[E^G : k] = 1$ and $k = E^G$.

(ii) ⇒ (iii) Let $p(x) \in k[x]$ be an irreducible polynomial having a root α in E, and let the distinct elements of the set $\{\sigma(\alpha) \colon \sigma \in G\}$ be $\alpha_1, \ldots, \alpha_n$. Define $g(x) \in E[x]$ by

$$g(x) = \prod (x - \alpha_i).$$

Now each $\sigma \in G$ permutes the α_i, so that each σ fixes each of the coefficients of $g(x)$; that is, the coefficients of $g(x)$ lie in $E^G = k$. Hence $g(x)$ is a polynomial in $k[x]$ having no repeated roots. Now $p(x)$ and $g(x)$ have a common root in E, and so their gcd in $E[x]$ is not 1; it follows from Corollary 3.41 that their gcd is not 1 in $k[x]$. Since $p(x)$ is irreducible, it must divide $g(x)$. Therefore, $p(x)$ has no repeated roots, hence is separable, and it splits over E.

(iii) ⇒ (i) Choose $\alpha_1 \in E$ with $\alpha_1 \notin k$. Since E/k is a finite extension, α_1 must be algebraic over k; let $p_1(x) = \text{irr}(\alpha_1, k) \in k[x]$ be its minimal polynomial. By hypothesis, $p_1(x)$ is a separable polynomial that splits over E; let $K_1 \subseteq E$ be its splitting field. If $K_1 = E$, we are done. Otherwise, choose $\alpha_2 \in E$ with $\alpha_2 \notin K_1$. By hypothesis, there is a separable irreducible $p_2(x) \in k[x]$ having α_2 as a root. Let $K_2 \subseteq E$ be the splitting field of $p_1(x)p_2(x)$, a separable polynomial. If $K_2 = E$, we are done; otherwise, repeat this construction. This process must end with $K_m = E$ for some m because E/k is finite. Thus, E is a splitting field of the separable polynomial $p_1(x) \cdots p_m(x)$. •

Definition. A field extension E/k is a ***Galois extension*** if it satisfies any of the equivalent conditions in Theorem 4.34.

Example 4.35.

If E/k is a finite separable extension, then the radical extension of E constructed in Lemma 4.17 is a Galois extension. ◂

Corollary 4.36. *If E/k is a Galois extension and if B is an **intermediate field**, that is, a subfield B with $k \subseteq B \subseteq E$, then E/B is a Galois extension.*

Proof. We know that E is a splitting field of some separable polynomial $f(x) \in k[x]$; that is, $E = k(\alpha_1, \dots, \alpha_n)$, where $\alpha_1, \dots, \alpha_n$ are the roots of $f(x)$. Since $k \subseteq B \subseteq E$, we have $f(x) \in B[x]$ and $E = B(\alpha_1, \dots, \alpha_n)$. •

Recall that the *elementary symmetric functions* of n variables are the polynomials, for $j = 1, \dots, n$,

$$e_j(x_1, \dots, x_n) = \sum_{i_1 < \cdots < i_j} x_{i_1} \cdots x_{i_j}.$$

If $z_1, \dots, z_n$ are the roots of $x^n + a_{n-1}x^{n-1} + \cdots + a_0$, then $e_j(z_1, \dots, z_n) = (-1)^j a_{n-j}$.

Theorem 4.37 (Fundamental Theorem of Symmetric Functions). *If k is a field, every symmetric function in $k(x_1, \dots, x_n)$ is a rational function in the elementary symmetric functions $e_1, \dots, e_n$.*

Proof. Let F be the smallest subfield of $E = k(x_1, \dots, x_n)$ containing the elementary symmetric functions. As we saw in Example 3.125, E is the splitting field of the general polynomial $f(t)$ of degree n:

$$f(t) = \prod_{i=1}^{n} (t - x_i).$$

As $f(t)$ is a separable polynomial, E/F is a Galois extension. We saw, in the proof of Theorem 4.27, the Abel–Ruffini theorem, that $\mathrm{Gal}(E/F) \cong S_n$. Therefore, $E^{S_n} = F$, by Theorem 4.34. But to say that $\theta(x) = g(x_1, \dots, x_n)/h(x_1, \dots, x_n)$ lies in E^{S_n} is to say that it is unchanged by permuting its variables; that is, $\theta(x)$ is a symmetric function. •

Exercise 6.84 on page 410 shows that every symmetric *polynomial* in $k[x_1, \dots, x_n]$ lies in $k[e_1, \dots, e_n]$.

Definition. If A and B are subfields of a field E, then their ***compositum***, denoted by $A \vee B$, is the intersection of all the subfields of E that contain $A \cup B$.

It is easy to see that $A \vee B$ is the smallest subfield of E containing both A and B. For example, if E/k is an extension with intermediate fields $A = k(\alpha_1, \dots, \alpha_n)$ and $B = k(\beta_1, \dots, \beta_m)$, then their compositum is

$$k(\alpha_1, \dots, \alpha_n) \vee k(\beta_1, \dots, \beta_m) = k(\alpha_1, \dots, \alpha_n, \beta_1, \dots, \beta_m).$$

Proposition 4.38.

(i) *Every Galois extension E/k is a separable extension of k.*

(ii) *If E/k is an algebraic field extension and $S \subseteq E$ is any, possibly infinite,*[7] *set of separable elements, then $k(S)/k$ is a separable extension.*

(iii) *Let E/k be an algebraic extension, where k is a field, and let B and C be intermediate fields. If both B/k and C/k are separable extensions, then their compositum $B \vee C$ is also a separable extension of k.*

Proof. (i) If $\beta \in E$, then $p(x) = \text{irr}(\beta, k) \in k[x]$ is an irreducible polynomial in $k[x]$ having a root in E. By Theorem 4.34(iii), $p(x)$ is a separable polynomial (which splits in $E[x]$). Therefore, β is separable over k, and E/k is a separable extension.

(ii) Let us first consider the case when S is finite; that is, $B = k(\alpha_1, \ldots, \alpha_t)$ is a finite extension, where each α_i is separable over k. By Lemma 4.17(i), there is an extension E/B that is a splitting field of some separable polynomial $f(x) \in k[x]$; hence, E/k is a Galois extension, by Theorem 4.34(i). By part (i) of this proposition, E/k is a separable extension; that is, for all $\alpha \in E$, the polynomial $\text{irr}(\alpha, k)$ has no repeated roots. In particular, $\text{irr}(\alpha, k)$ has no repeated roots for all $\alpha \in B$, and so B/k is a separable extension.

We now consider the general case. If $\alpha \in k(S)$, then Exercise 3.95 on page 197 says that there are finitely many elements $\alpha_1, \ldots, \alpha_n \in S$ with $\alpha \in B = k(\alpha_1, \ldots, \alpha_n)$. As we have just seen, B/k is a separable extension, and so α is separable over k. As α is an arbitrary element of $k(S)$, it follows that $k(S)/k$ is a separable extension.

(iii) Apply part (i) to the subset $S = B \cup C$, for $B \vee C = k(B \cup C)$. •

Query: If E/k is a Galois extension and B is an intermediate field, is B/k a Galois extension? The answer is no; in Example 4.29, we saw that $E = \mathbb{Q}(\sqrt[3]{2}, \omega)$ is a splitting field of $x^3 - 2$ over $\mathbb{Q}$, where ω is a primitive cube root of unity, and so it is a Galois extension. However, the intermediate field $B = \mathbb{Q}(\sqrt[3]{2})$ is not a Galois extension, for $x^3 - 2$ is an irreducible polynomial having a root in B, yet it does not split in $B[x]$.

The following proposition determines when an intermediate field B does give a Galois extension.

Definition. If E/k is a Galois extension and if B is an intermediate field, then a ***conjugate*** of B is an intermediate field of the form

$$B^\sigma = \{\sigma(b) : b \in B\}$$

for some $\sigma \in \text{Gal}(E/k)$.

[7]This result is true if finitely many transcendental elements are adjoined (remember that transcendental elements are always separable, by definition), but it may be false if infinitely many transcendental elements are adjoined.

Proposition 4.39. *If E/k is a Galois extension, then an intermediate field B has no conjugates other than B itself if and only if B/k is a Galois extension.*

Proof. Assume that $B^\sigma = B$ for all $\sigma \in G$, where $G = \mathrm{Gal}(E/k)$. Let $p(x) \in k[x]$ be an irreducible polynomial having a root β in B. Since $B \subseteq E$ and E/k is Galois, $p(x)$ is a separable polynomial and it splits in $E[x]$. If $\beta' \in E$ is another root of $p(x)$, there exists an isomorphism $\sigma \in G$ with $\sigma(\beta) = \beta'$ (for G acts transitively on the roots of an irreducible polynomial, by Proposition 4.13). Therefore, $\beta' = \sigma(\beta) \in B^\sigma = B$, so that $p(x)$ splits in $B[x]$. Therefore, B/k is a Galois extension.

Conversely, since B/k is a splitting field of some polynomial $f(x)$ over k, we have $B = k(\alpha_1, \ldots, \alpha_n)$, where $\alpha_1, \ldots, \alpha_n$ are all the roots of $f(x)$. Since every $\sigma \in \mathrm{Gal}(E/k)$ must permute the roots of $f(x)$, it follows that σ must send B to itself. •

We are now going to show, when E/k is a Galois extension, that the intermediate fields are classified by the subgroups of $\mathrm{Gal}(E/k)$.

We begin with some general definitions.

Definition. A set X is a ***partially ordered set*** if it has a binary relation $x \preceq y$ defined on it that satisfies, for all $x, y, z \in X$,

(i) ***Reflexivity***: $x \preceq x$;

(ii) ***Antisymmetry***: If $x \preceq y$, and $y \preceq x$, then $x = y$;

(iii) ***Transitivity***: If $x \preceq y$ and $y \preceq z$, then $x \preceq z$.

An element c in a partially ordered set X is an ***upper bound*** of $a, b \in X$ if $a \preceq c$ and $b \preceq c$; an element $d \in X$ is a ***least upper bound*** of a, b if d is an upper bound and if $d \preceq c$ for every upper bound c of a and b. ***Lower bounds*** and ***greatest lower bounds*** are defined similarly, everywhere reversing the inequalities.

We will discuss partially ordered sets more thoroughly in the Appendix. Here, we are more interested in special partially ordered sets called *lattices*.

Definition. A ***lattice*** is a partially ordered set $\mathcal{L}$ in which every pair of elements $a, b \in \mathcal{L}$ has a greatest lower bound $a \wedge b$ and a least upper bound $a \vee b$.

Example 4.40.

(i) If U is a set, define $\mathcal{L}$ to be the family of all the subsets of U, and define $A \preceq B$ to mean $A \subseteq B$. Then $\mathcal{L}$ is a lattice, where $A \wedge B = A \cap B$ and $A \vee B = A \cup B$.

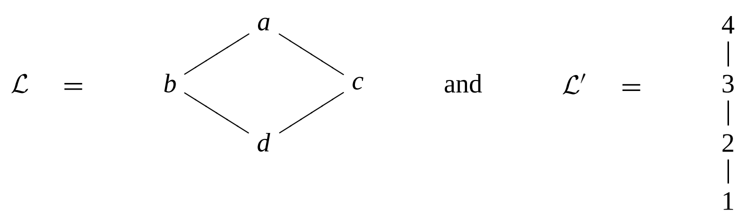
(ii) If G is a group, define $\mathcal{L} = \mathrm{Sub}(G)$ to be the family of all the subgroups of G, and define $A \preceq B$ to mean $A \leq B$; that is, A is a subgroup of B. Then $\mathcal{L}$ is a lattice, where $A \wedge B = A \cap B$ and $A \vee B$ is the subgroup generated by $A \cup B$.

(iii) If E/k is a field extension, define $\mathcal{L} = \mathrm{Int}(E/k)$ to be the family of all the intermediate fields, and define $K \preceq B$ to mean $K \subseteq B$; that is, K is a subfield of B. Then $\mathcal{L}$ is a lattice, where $K \wedge B = K \cap B$ and $K \vee B$ is the compositum of K and B.

(iv) If n is a positive integer, define $\mathrm{Div}(n)$ to be the set of all the positive divisors of n. Then $\mathrm{Div}(n)$ is a partially ordered set if one defines $d \preceq d'$ to mean $d \mid d'$. Here, $d \wedge d' = \gcd(d, d')$ and $d \vee d' = \mathrm{lcm}(d, d')$. ◂

Definition. If $\mathcal{L}$ and $\mathcal{L}'$ are lattices, a function $f : \mathcal{L} \to \mathcal{L}'$ is called ***order-reversing*** if $a \preceq b$ in $\mathcal{L}$ implies $f(b) \preceq f(a)$ in $\mathcal{L}'$.

Example 4.41.
There exist lattices $\mathcal{L}$ and $\mathcal{L}'$ and an order-reversing bijection $\varphi\colon \mathcal{L} \to \mathcal{L}'$ whose inverse $\varphi^{-1}\colon \mathcal{L}' \to \mathcal{L}$ is not order-reversing. For example, consider the lattices



The bijection $\varphi\colon \mathcal{L} \to \mathcal{L}'$, defined by

$$\varphi(a) = 1, \quad \varphi(b) = 2, \quad \varphi(c) = 3, \quad \varphi(d) = 4,$$

is an order-reversing bijection, but its inverse $\varphi^{-1}\colon \mathcal{L}' \to \mathcal{L}$ is not order-reversing, because $2 \preceq 3$ but $c = \varphi^{-1}(3) \not\preceq \varphi^{-1}(2) = b$. ◂

The De Morgan laws say that if A and B are subsets of a set X, and if A' denotes the complement of A, then

$$(A \cap B)' = A' \cup B' \quad \text{and} \quad (A \cup B)' = A' \cap B'.$$

These identities are generalized in the next lemma.

Lemma 4.42. *Let $\mathcal{L}$ and $\mathcal{L}'$ be lattices, and let $\varphi\colon \mathcal{L} \to \mathcal{L}'$ be a bijection such that both φ and φ^{-1} are order-reversing. Then*

$$\varphi(a \wedge b) = \varphi(a) \vee \varphi(b) \quad \textit{and} \quad \varphi(a \vee b) = \varphi(a) \wedge \varphi(b).$$

Proof. Since $a, b \preceq a \vee b$, we have $\varphi(a \vee b) \preceq \varphi(a), \varphi(b)$; that is, $\varphi(a \vee b)$ is a lower bound of $\varphi(a), \varphi(b)$. It follows that $\varphi(a \vee b) \preceq \varphi(a) \wedge \varphi(b)$.

For the reverse inequality, surjectivity of φ gives $c \in \mathcal{L}$ with $\varphi(a) \wedge \varphi(b) = \varphi(c)$. Now $\varphi(c) = \varphi(a) \wedge \varphi(b) \preceq \varphi(a), \varphi(b)$. Applying φ^{-1}, which is also order-reversing, we have $a, b \preceq c$. Hence, c is an upper bound of a, b, so that $a \vee b \preceq c$. Therefore, $\varphi(a \vee b) \succeq \varphi(c) = \varphi(a) \wedge \varphi(b)$. A similar argument proves the other half of the statement. •

Theorem 4.43 (Fundamental Theorem of Galois Theory). *Let E/k be a finite Galois extension with Galois group $G = \mathrm{Gal}(E/k)$.*

(i) *The function $\gamma\colon \mathrm{Sub}(\mathrm{Gal}(E/k)) \to \mathrm{Int}(E/k)$, defined by*

$$\gamma\colon H \mapsto E^H,$$

is an order-reversing bijection whose inverse, $\delta\colon \mathrm{Int}(E/k) \to \mathrm{Sub}(\mathrm{Gal}(E/k))$, is the order-reversing bijection

$$\delta\colon B \mapsto \mathrm{Gal}(E/B).$$

(ii) *For every $B \in \mathrm{Int}(E/k)$ and $H \in \mathrm{Sub}(\mathrm{Gal}(E/k))$,*

$$E^{\mathrm{Gal}(E/B)} = B \quad \text{and} \quad \mathrm{Gal}(E/E^H) = H.$$

(iii) *For every $H, K \in \mathrm{Sub}(\mathrm{Gal}(E/k))$ and $B, C \in \mathrm{Int}(E/k)$,*

$$E^{H \vee K} = E^H \cap E^K;$$
$$E^{H \cap K} = E^H \vee E^K;$$
$$\mathrm{Gal}(E/(B \vee C)) = \mathrm{Gal}(E/B) \cap \mathrm{Gal}(E/C);$$
$$\mathrm{Gal}(E/(B \cap C)) = \mathrm{Gal}(E/B) \vee \mathrm{Gal}(E/C).$$

(iv) *For every $B \in \mathrm{Int}(E/k)$ and $H \in \mathrm{Sub}(\mathrm{Gal}(E/k))$,*

$$[B : k] = [G : \mathrm{Gal}(E/B)] \quad \text{and} \quad [G : H] = [E^H : k].$$

(v) *If $B \in \mathrm{Int}(E/k)$, then B/k is a Galois extension if and only if $\mathrm{Gal}(E/B)$ is a normal subgroup of G.*

Proof. (i) Proposition 4.28 proves that γ is order-reversing, and it is also easy to prove that δ is order-reversing. Now injectivity of γ is proved in Theorem 4.33, so that Proposition 1.47 shows that it suffices to prove that $\gamma\delta\colon \mathrm{Int}(E/k) \to \mathrm{Int}(E/k)$ is the identity; it will follow that γ is a bijection with inverse δ. If B is an intermediate field, then $\delta\gamma\colon B \mapsto E^{\mathrm{Gal}(E/B)}$. But E/E^B is a Galois extension, by Corollary 4.36, and so $E^{\mathrm{Gal}(E/B)} = B$, by Theorem 4.34.

(ii) This is just the statement that $\gamma\delta$ and $\delta\gamma$ are identity functions.

(iii) These statements follow from Lemma 4.42.

(iv) By Theorem 4.7(ii) and the fact that E/B is a Galois extension,

$$[B : k] = [E : k]/[E : B] = |G|/|\mathrm{Gal}(E/B)| = [G : \mathrm{Gal}(E/B)].$$

Thus, the degree of B/k is the index of its Galois group in G. The second equation follows from this one; take $B = E^H$, noting that (ii) gives $\mathrm{Gal}(E/E^H) = H$:

$$[E^H : k] = [G : \mathrm{Gal}(E/E^H)] = [G : H].$$

(v) It follows from Theorem 4.16 that $\mathrm{Gal}(E/B) \lhd G$ when B/k is a Galois extension (both B/k and E/k are splitting fields of polynomials in $k[x]$). For the converse, let $H = \mathrm{Gal}(E/B)$, and assume that $H \lhd G$. Now $E^H = E^{\mathrm{Gal}(E/B)} = B$, by (ii), and so it suffices to prove that $(E^H)^\sigma = E^H$ for every $\sigma \in G$, by Proposition 4.39. Suppose now that $a \in E^H$; that is, $\eta(a) = a$ for all $\eta \in H$. If $\sigma \in G$, then we must show that $\eta(\sigma(a)) = \sigma(a)$ for all $\eta \in H$. Now $H \lhd G$ says that if $\eta \in H$ and $\sigma \in G$, then there is $\eta' \in H$ with $\eta\sigma = \sigma\eta'$ (of course, $\eta' = \sigma^{-1}\eta\sigma$). But

$$\eta\sigma(a) = \sigma\eta'(a) = \sigma(a),$$

because $\eta'(a) = a$, as desired. Therefore, $B/k = E^H/k$ is Galois. •

Here are some corollaries.

Theorem 4.44. *If E/k is a Galois extension whose Galois group is abelian, then every intermediate field is a Galois extension.*

Proof. Every subgroup of an abelian group is a normal subgroup. •

Corollary 4.45. *A Galois extension E/k has only finitely many intermediate fields.*

Proof. The finite group $\mathrm{Gal}(E/k)$ has only finitely many subgroups. •

Definition. A field extension E/k is a ***simple extension*** if there is $u \in E$ with $E = k(u)$.

The following theorem of E. Steinitz characterizes simple extensions.

Theorem 4.46 (Steinitz). *A finite extension E/k is simple if and only if it has only finitely many intermediate fields.*

Proof. Assume that E/k is a simple extension, so that $E = k(u)$; let $p(x) = \mathrm{irr}(u, k) \in k[x]$ be its minimal polynomial. If B is any intermediate field, let

$$q(x) = \mathrm{irr}(u, B) = b_0 + b_1 x + \cdots + b_{n-1}x^{n-1} + x^n \in B[x]$$

be the monic irreducible polynomial of u over B, and define

$$B' = k(b_0, \ldots, b_{n-1}) \subseteq B.$$

Note that $q(x)$ is an irreducible polynomial over the smaller field B'. Now

$$E = k(u) \subseteq B'(u) \subseteq B(u) \subseteq E,$$

so that $B'(u) = E = B(u)$. Hence, $[E : B] = [B(u) : B]$ and $[E : B'] = [B'(u) : B']$. But each of these is equal to $\deg(q)$, by Proposition 3.117(v), so that $[E : B] = \deg(q) = [E : B']$. Since $B' \subseteq B$, it follows that $[B : B'] = 1$; that is,

$$B = B' = k(b_0, \ldots, b_{n-1}).$$

We have characterized B in terms of the coefficients of $q(x)$, a monic divisor of $p(x) = \operatorname{irr}(u, k)$ in $E[x]$. But $p(x)$ has only finitely many monic divisors, and hence there are only finitely many intermediate fields.

Conversely, assume that E/k has only finitely many intermediate fields. If k is a finite field, then we know that E/k is a simple extension (take u to be a primitive element); therefore, we may assume that k is infinite. Since E/k is a finite extension, there are elements $u_1, \ldots, u_n$ with $E = k(u_1, \ldots, u_n)$. By induction on $n \geq 1$, it suffices to prove that $E = k(a, b)$ is a simple extension. Now there are infinitely many elements $c \in E$ of the form $c = a + tb$, where $t \in k$, for k is now infinite. Since there are only finitely many intermediate fields, there are, in particular, only finitely many fields of the form $k(c)$. By the pigeonhole principle,[8] there exist distinct elements $t, t' \in k$ with $k(c) = k(c')$, where $c' = a + t'b$. Clearly, $k(c) \subseteq k(a, b)$. For the reverse inclusion, the field $k(c) = k(c')$ contains $c - c' = (t - t')b$, so that $b \in k(c)$ (because $t - t' \neq 0$). It follows that $a = c - tb \in k(c)$, and so $k(c) = k(a, b)$. •

An immediate consequence is that every Galois extension is simple; in fact, even more is true.

Theorem 4.47 (Theorem of the Primitive Element). *If B/k is a finite separable extension, then there is $u \in B$ with $B = k(u)$. In particular, if k has characteristic 0, then every finite extension B/k is a simple extension.*

Proof. By Example 4.35, the radical extension E/k constructed in Lemma 4.17 is a Galois extension having B as an intermediate field, so that Corollary 4.45 says that the extension E/k has only finitely many intermediate fields. It follows at once that the extension B/k has only finitely many intermediate fields, and so Steinitz's theorem says that B/k has a primitive element. •

The theorem of the primitive element was known by Lagrange, and Galois used a modification of it in order to construct the original version of the Galois group.

We now turn to finite fields.

Theorem 4.48. *The finite field $\mathbb{F}_q$, where $q = p^n$, has exactly one subfield of order p^d for every divisor d of n, and no others.*

Proof. First, $\mathbb{F}_q/\mathbb{F}_p$ is a Galois extension, for it is a splitting field of the separable polynomial $x^q - x$. Now $G = \operatorname{Gal}(\mathbb{F}_q/\mathbb{F}_p)$ is cyclic of order n, by Theorem 4.12. Since a cyclic group of order n has exactly one subgroup of order d for every divisor d of n, by Lemma 2.85, it follows that G has exactly one subgroup H of index n/d. Therefore, there is only one intermediate field, namely, E^H, with $[E^H : \mathbb{F}_p] = [G : H] = n/d$, and $E^H = \mathbb{F}_{p^{n/d}}$. •

We now give two algebraic proofs of the fundamental theorem of algebra, proved by Gauss (1799): The first, due to P. Samuel (which he says is "by a method essentially due

[8] If there is an infinite number of pigeons in only finitely many pigeonholes, then at least one of the holes contains an infinite number of pigeons.

to Lagrange"), uses the fundamental theorem of symmetric functions; the second uses the fundamental theorem of Galois theory, as well as a Sylow theorem which we will prove in Chapter 5.

Assume that $\mathbb{R}$ satisfies a weak form of the intermediate value theorem: If $f(x) \in \mathbb{R}[x]$ and there exist $a, b \in \mathbb{R}$ such that $f(a) > 0$ and $f(b) < 0$, then $f(x)$ has a real root. Here are some preliminary consequences.

(i) *Every positive real number r has a real square root.*

If $f(x) = x^2 - r$, then

$$f(1+r) = (1+r)^2 - r = 1 + r + r^2 > 0,$$

and $f(0) = -r < 0$.

(ii) *Every quadratic $g(x) \in \mathbb{C}[x]$ has a complex root.*

First, every complex number z has a complex square root: When z is written in polar form $z = re^{i\theta}$, where $r \geq 0$, then $\sqrt{z} = \sqrt{r}e^{i\theta/2}$. The quadratic formula gives the (complex) roots of $g(x)$.

(iii) *The field $\mathbb{C}$ has no extensions of degree* 2.

Such an extension would contain an element whose minimal polynomial is an irreducible quadratic in $\mathbb{C}[x]$; but Item (ii) shows that no such polynomial exists.

(iv) *Every $f(x) \in \mathbb{R}[x]$ having odd degree has a real root.*

Let $f(x) = a_0 + a_1x + \cdots + a_{n-1}x^{n-1} + x^n \in \mathbb{R}[x]$. Define $t = 1 + \sum |a_i|$. Now $|a_i| \leq t - 1$ for all i and, if $h(x) = f(x) - x^n$, then

$$\begin{aligned} |h(t)| &= \left|a_0 + a_1t + \cdots + a_{n-1}t^{n-1}\right| \\ &\leq (t-1)\left(1 + t + \ldots + t^{n-1}\right) \\ &= t^n - 1 \\ &< t^n. \end{aligned}$$

Therefore, $-t^n < h(t)$ and $0 = -t^n + t^n < h(t) + t^n = f(t)$.

A similar argument shows that $|h(-t)| < t^n$, so that

$$f(-t) = h(-t) + (-t)^n < t^n + (-t)^n.$$

When n is odd, $(-t)^n = -t^n$, and so $f(-t) < t^n - t^n = 0$. Therefore, the intermediate value theorem provides a real number r with $f(r) = 0$; that is, $f(x)$ has a real root.

(v) *There is no field extension $E/\mathbb{R}$ of odd degree* > 1.

If $u \in E$, then its minimal polynomial $\operatorname{irr}(u, \mathbb{R})$ must have even degree, by Item (iv), so that $[\mathbb{R}(u) : \mathbb{R}]$ is even. Hence $[E : \mathbb{R}] = [E : \mathbb{R}(u)][\mathbb{R}(u) : \mathbb{R}]$ is even.

Theorem 4.49 (Fundamental Theorem of Algebra). *If $f(x) \in \mathbb{C}[x]$ has degree $n \geq 1$, then $f(x)$ has a complex root, and hence $f(x)$ splits: There are $c, u_1, \ldots, u_n \in \mathbb{C}$ with*

$$f(x) = c(x - u_1) \cdots (x - u_n).$$

Proof. We show that $f(x) = \sum a_i x^i \in \mathbb{C}[x]$ has a complex root. Define $\overline{f}(x) = \sum \overline{a}_i x^i$, where $\overline{a}_i$ is the complex conjugate of a_i. Now $f(x)\overline{f}(x) = \sum c_k x^k$, where $c_k = \sum_{i+j=k} a_i \overline{a}_j$; hence, $\overline{c}_k = c_k$, so that $f(x)\overline{f}(x) \in \mathbb{R}[x]$. If $f(x)$ has a complex root z, then z is a root of $f(x)\overline{f}(x)$. Conversely, if z is a complex root of $f(x)\overline{f}(x)$, then z is a root of either $f(x)$ or $\overline{f}(x)$. But if z is a root of $\overline{f}(x)$, then $\overline{z}$ is a root of $f(x)$. Therefore, $f(x)$ has a complex root if and only if $f(x)\overline{f}(x)$ has a complex root, and so it suffices to prove that every real polynomial has a complex root.

To summarize, it suffices to prove that every nonconstant monic $f(x) \in \mathbb{R}[x]$ has a complex root. Let $\deg(f) = 2^k m$, where m is odd; we prove the result by induction on $k \geq 0$. The base step $k = 0$ is proved in Item (iv), and so we may assume that $k \geq 1$. Let $\alpha_1, \ldots, \alpha_n$ be the roots of $f(x)$ in some splitting field of $f(x)$. For fixed $t \in \mathbb{R}$, define

$$g_t(x) = \prod_{\{i,j\}} (x - \beta_{ij}),$$

where $\beta_{ij} = \alpha_i + \alpha_j + t\alpha_i\alpha_j$ and $\{i, j\}$ varies over all the two-element subsets of $\{1, \ldots, n\}$. First,

$$\deg(g_t) = \tfrac{1}{2}n(n-1) = 2^{k-1}m(n-1).$$

Now $n = 2^k m$ is even, because $k \geq 1$, so that $n - 1$ is odd; hence, $m(n-1)$ is odd. Thus, the inductive hypothesis will apply if $g_t(x) \in \mathbb{R}[x]$.

For each coefficient c of $g_t(x)$, there is an elementary symmetric function

$$e(\ldots, y_{ij}, \ldots) \in \mathbb{R}[\ldots, y_{ij}, \ldots]$$

with $c = e(\ldots, \beta_{ij}, \ldots)$. If we define

$$h(x_1, \ldots, x_n) = e(\ldots, x_i + x_j + tx_ix_j, \ldots),$$

then

$$c = e(\ldots, \alpha_i + \alpha_j + t\alpha_i\alpha_j, \ldots) = h(\alpha_1, \ldots, \alpha_n).$$

Each $\sigma \in S_n$ acts on $\mathbb{R}[x_1, \ldots, x_n]$ via $\sigma\colon x_i + x_j + tx_ix_j \mapsto x_{\sigma i} + x_{\sigma j} + tx_{\sigma i}x_{\sigma j}$, and hence it permutes the set of polynomials of this form. Since the elementary symmetric function $e(\ldots, y_{ij}, \ldots)$ is invariant under every permutation of the variables y_{ij}, it follows that $h(x_1, \ldots, x_n) = E(\ldots, x_i + x_j + tx_ix_j, \ldots)$ is a symmetric function of $x_1, \ldots, x_n$. By the fundamental theorem of symmetric polynomials (Exercise 6.84 on page 410), there is a polynomial $\varphi(x) \in \mathbb{R}[x_1, \ldots, x_n]$ with

$$h(x_1, \ldots, x_n) = \varphi(e_1(x_1, \ldots, x_n), \ldots, e_n(x_1, \ldots, x_n)).$$

The evaluation $(x_1, \ldots, x_n) \mapsto (\alpha_1, \ldots, \alpha_n)$ gives

$$c = h(\alpha_1, \ldots, \alpha_n) = \varphi(e_1(\alpha_1, \ldots, \alpha_n), \ldots, e_n(\alpha_1, \ldots, \alpha_n)).$$

But $e_r(\alpha_1, \ldots, \alpha_n)$ is just the rth coefficient of $f(x)$, which is real, and so c is real; that is, $g_t(x) \in \mathbb{R}[x]$.

By induction, $g_t(x)$ has a complex root for each $t \in \mathbb{R}$. There are infinitely many $t \in \mathbb{R}$ and only finitely many two-element subsets $\{i, j\}$. By the pigeonhole principle, there exists a subset $\{i, j\}$ and distinct reals t and s with both $\alpha_i + \alpha_j + t\alpha_i\alpha_j$ and $\alpha_i + \alpha_j + s\alpha_i\alpha_j$ complex [for the β_{ij} are the roots of $g_t(x)$]. Subtracting, $(t - s)\alpha_i\alpha_j \in \mathbb{C}$; as $t \neq s$, we have $\alpha_i\alpha_j \in \mathbb{C}$; say, $\alpha_i\alpha_j = u$. Since $\alpha_i + \alpha_j + t\alpha_i\alpha_j \in \mathbb{C}$, it follows that $\alpha_i + \alpha_j \in \mathbb{C}$; say, $\alpha_i + \alpha_j = v$. Therefore, α_i is a root of $x^2 - vx + u$, and the quadratic formula, Item (ii), gives $\alpha_i \in \mathbb{C}$, as desired. That $f(x)$ splits now follows by induction on $n \geq 1$. •

Here is a second proof.

Theorem (Fundamental Theorem of Algebra). *Every nonconstant $f(x) \in \mathbb{C}[x]$ has a complex root.*

Proof. As in the proof just given, it suffices to prove that every nonconstant $f(x) \in \mathbb{R}[x]$ has a complex root. Let $E/\mathbb{R}$ be a splitting field of $(x^2 + 1)f(x)$ that contains $\mathbb{C}$. Since $\mathbb{R}$ has characteristic 0, $E/\mathbb{R}$ is a Galois extension; let $G = \mathrm{Gal}(E/\mathbb{R})$ be its Galois group. Now $|G| = 2^m k$, where $m \geq 0$ and k is odd. By the Sylow theorem (Theorem 5.36), G has a subgroup H of order 2^m; let $B = E^H$ be the corresponding intermediate field. By the fundamental theorem of Galois theory, the degree $[B : \mathbb{R}]$ is equal to the index $[G : H] = k$. But we have seen, in Item (v), that $\mathbb{R}$ has no extension of odd degree greater than 1; hence $k = 1$ and G is a 2-group. Now $E/\mathbb{C}$ is also a Galois extension, and $\mathrm{Gal}(E/\mathbb{C}) \leq G$ is also a 2-group. If this group is nontrivial, then it has a subgroup K of index 2. By the fundamental theorem once again, the intermediate field E^K is an extension of $\mathbb{C}$ of degree 2, and this contradicts Item (iii). We conclude that $[E : \mathbb{C}] = 1$; that is, $E = \mathbb{C}$. But E is a splitting field of $f(x)$ over $\mathbb{C}$, and so $f(x)$ has a complex root. •

We now prove the converse of Galois's theorem (which holds only in characteristic 0): Solvability of the Galois group implies solvability by radicals of the polynomial. It will be necessary to prove that a certain field extension is a pure extension, and we will use the *norm* (which arises quite naturally in algebraic number theory; for example, it was used in the proof of Theorem 3.66, Fermat's two-squares theorem).

Definition. If E/k is a Galois extension and $u \in E^\times$, define its ***norm*** $N(u)$ by

$$N(u) = \prod_{\sigma \in \mathrm{Gal}(E/k)} \sigma(u).$$

Here are some preliminary properties of the norm, whose simple proofs are left as exercises.

(i) If $u \in E^\times$, then $N(u) \in k^\times$ (because $N(u) \in E^G = k$).

(ii) $N(uv) = N(u)N(v)$, so that $N\colon E^\times \to k^\times$ is a homomorphism.

(iii) If $a \in k$, then $N(a) = a^n$, where $n = [E : k]$.

(iv) If $\sigma \in G$ and $u \in E^\times$, then $N(\sigma(u)) = N(u)$.

Given a homomorphism, we ask about its kernel and image. The image of the norm is not easy to compute; the next result (which was the ninetieth theorem in an 1897 exposition of Hilbert on algebraic number theory) computes the kernel of the norm in a special case.

Theorem 4.50 (Hilbert's Theorem 90). *Let E/k be a Galois extension whose Galois group $G = \mathrm{Gal}(E/k)$ is cyclic of order n, say, with generator σ. If $u \in E^\times$, then $N(u) = 1$ if and only if there exists $v \in E^\times$ with $u = v\sigma(v)^{-1}$.*

Proof. If $u = v\sigma(v)^{-1}$, then

$$\begin{aligned} N(u) &= N(v\sigma(v)^{-1}) \\ &= N(v)N(\sigma(v)^{-1}) \\ &= N(v)N(\sigma(v))^{-1} \\ &= N(v)N(v)^{-1} = 1. \end{aligned}$$

Conversely, let $N(u) = 1$. Define "partial norms" in $E^\times$:

$$\begin{aligned} \delta_0 &= u, \\ \delta_1 &= u\sigma(u), \\ \delta_2 &= u\sigma(u)\sigma^2(u), \\ &\vdots \\ \delta_{n-1} &= u\sigma(u)\cdots\sigma^{n-1}(u). \end{aligned}$$

Note that $\delta_{n-1} = N(u) = 1$. It is easy to see that

$$u\sigma(\delta_i) = \delta_{i+1} \text{ for all } 0 \le i \le n-2. \tag{4}$$

By independence of the characters $1, \sigma, \sigma^2, \ldots, \sigma^{n-1}$, there exists $y \in E$ with

$$\delta_0 y + \delta_1\sigma(y) + \cdots + \delta_{n-2}\sigma^{n-2}(y) + \sigma^{n-1}(y) \neq 0;$$

call this sum z. Using Eq. (4), we easily check that

$$\begin{aligned} \sigma(z) &= \sigma(\delta_0)\sigma(y) + \sigma(\delta_1)\sigma^2(y) + \cdots + \sigma(\delta_{n-2})\sigma^{n-1}(y) + \sigma^n(y) \\ &= u^{-1}\delta_1\sigma(y) + u^{-1}\delta_2\sigma^2(y) + \cdots + u^{-1}\delta_{n-1}\sigma^{n-1}(y) + y \\ &= u^{-1}\Big(\delta_1\sigma(y) + \delta_2\sigma^2(y) + \cdots + \delta_{n-1}\sigma^{n-1}(y)\Big) + u^{-1}\delta_0 y \\ &= u^{-1}z. \quad \bullet \end{aligned}$$

Corollary 4.51. *Let E/k be a Galois extension of prime degree p. If k contains a primitive pth root of unity ω, then $E = k(z)$, where $z^p \in k$, and so E/k is a pure extension of type p.*

Proof. The Galois group $G = \mathrm{Gal}(E/k)$ has order p, hence is cyclic; let σ be a generator. Observe that $N(\omega) = \omega^p = 1$, because $\omega \in k$. By Hilbert's Theorem 90, we have $\omega = z\sigma(z)^{-1}$ for some $z \in E$. Hence $\sigma(z) = \omega^{-1}z$. Thus, $\sigma(z^p) = (\omega^{-1}z)^p = z^p$, and so $z^p \in E^G$, because σ generates G; since E/k is Galois, however, we have $E^G = k$, so that $z^p \in k$. Note that $z \notin k$, lest $\omega = 1$, so that $k(z) \neq k$ is an intermediate field. Therefore $E = k(z)$, because $[E : k] = p$ is prime, and hence E has no proper intermediate fields. •

We confess that we have presented Hilbert's Theorem 90, not only because of its corollary, which will be used to prove Galois's theorem, but also because it is a well-known result that is an early instance of homological algebra (see Corollary 10.129). Here is an elegant proof of Corollary 4.51 due to E. Houston (we warn the reader that it uses eigenvalues, a topic we have not yet introduced).

Proposition 4.52. *Let E/k be a Galois extension of prime degree p. If k contains a primitive pth root of unity ω, then $E = k(z)$, where $z^p \in k$, and so E/k is a pure extension of type p.*

Proof. Since E/k is a Galois extension of degree p, its Galois group $G = \mathrm{Gal}(E/k)$ has order p, and hence it is cyclic: $G = \langle\sigma\rangle$. View E as a vector space over k. If $a \in k$ and $u \in E$, then $\sigma(au) = \sigma(a)\sigma(u) = a\sigma(u)$, because $\sigma \in \mathrm{Gal}(E/k)$ (so that it fixes k), and so we may view $\sigma \colon E \to E$ as a linear transformation. Now σ satisfies the polynomial $x^p - 1$, because $\sigma^p = 1_E$, by Lagrange's theorem. But σ satisfies no polynomial of smaller degree, lest we contradict independence of the characters $1, \sigma, \sigma^2, \ldots, \sigma^{p-1}$. Therefore, $x^p - 1$ is the minimum polynomial of σ, and so every pth root of unity ω is an eigenvalue of σ. Since $\omega^{-1} \in k$, by hypothesis, there is some eigenvector $z \in E$ of σ with $\sigma(z) = \omega^{-1}z$ (note that $z \notin k$ because it is not fixed by σ). Hence, $\sigma(z^p) = (\sigma(z))^p = (\omega^{-1})^p z^p = z^p$, from which it follows that $z^p \in E^G = k$. Now $p = [E : k] = [E : k(z)][k(z) : k]$; since p is prime and $[k(z) : k] \neq 1$, we have $[E : k(z)] = 1$; that is, $E = k(z)$, and so E/k is a pure extension. •

Theorem 4.53 (Galois). *Let k be a field of characteristic 0, let E/k be a Galois extension, and let $G = \mathrm{Gal}(E/k)$ be a solvable group. Then E can be imbedded in a radical extension of k.*

Therefore, the Galois group of a polynomial over a field of characteristic 0 is a solvable group if and only if the polynomial is solvable by radicals.

Remark. A counterexample in characteristic p is given in Proposition 4.56. ◂

Proof. Since G is solvable, it has a normal subgroup H of prime index, say, p. Let ω be a primitive pth root of unity, which exists in some extension field, because k has characteristic 0. We distinguish two cases.

Case (i): $\omega \in k$.

We prove the statement by induction on $[E : k]$. The base step is obviously true, for $k = E$ is a radical extension of itself. For the inductive step, consider the intermediate field E^H. Now E/E^H is a Galois extension, by Corollary 4.36, and $\mathrm{Gal}(E/E^H)$ is solvable, being a subgroup of the solvable group G. Since $[E : E^H] < [E : k]$, the inductive hypothesis gives a radical tower $E^H \subseteq R_1 \subseteq \cdots \subseteq R_t$, where $E \subseteq R_t$. Now E^H/k is a Galois extension, because $H \lhd G$, and its index $[G : H] = p = [E^H : k]$, by the fundamental theorem. Corollary 4.51 (or Proposition 4.52) now applies to give $E^H = k(z)$, where $z^p \in k$; that is, E^H/k is a pure extension. Hence, the radical tower above can be lengthened by adding the prefix $k \subseteq E^H$, thus displaying R_t/k as a radical extension.

Case (ii): General case.

Let $k^* = k(\omega)$, and define $E^* = E(\omega)$. We claim that E^*/k is a Galois extension. Since E/k is a Galois extension, it is the splitting field of some separable $f(x) \in k[x]$, and so E^* is a splitting field over k of $f(x)(x^p - 1)$. But $x^p - 1$ is separable, because k has characteristic 0, and so E^*/k is a Galois extension. Therefore, E^*/k^* is also a Galois extension, by Corollary 4.36. Let $G^* = \mathrm{Gal}(E^*/k^*)$. By Exercise 4.5 on page 217, accessory irrationalities, there is an injection $\psi\colon G^* \to G = \mathrm{Gal}(E/k)$, so that G^* is solvable, being isomorphic to a subgroup of a solvable group. Since $\omega \in k^*$, the first case says that there is a radical tower $k^* \subseteq R_1^* \subseteq \cdots \subseteq R_m^*$ with $E \subseteq E^* \subseteq R_m^*$. But $k^* = k(\omega)$ is a pure extension, so that this last radical tower can be lengthened by adding the prefix $k \subseteq k^*$, thus displaying R_m^*/k as a radical extension. •

We now have another proof of the existence of the classical formulas.

Corollary 4.54. *If k has characteristic 0, then every $f(x) \in k[x]$ with $\deg(f) \leq 4$ is solvable by radicals.*

Proof. If G is the Galois group of $f(x)$, then G is isomorphic to a subgroup of S_4. But S_4 is a solvable group, and so every subgroup of S_4 is also solvable. By Galois's theorem, $f(x)$ is solvable by radicals. •

Suppose we know the Galois group G of a polynomial $f(x) \in \mathbb{Q}[x]$ and that G is solvable. Can we use this information to find the roots of $f(x)$? The answer is affirmative; we suggest the reader look at the book by Gaal, *Classical Galois Theory with Examples*, to see how this is done.

In 1827, N. H. Abel proved that if the Galois group of a polynomomial $f(x)$ is commutative, then $f(x)$ is solvable by radicals (of course, Galois groups had not yet been defined). This result was superseded by Galois's theorem, proved in 1830, but it is the reason why abelian groups are so called.

A deep theorem of W. Feit and J. G. Thompson (1963) says that every group of odd order is solvable. It follows that if k is a field of characteristic 0 and $f(x) \in k[x]$ is a polynomial whose Galois group has odd order, equivalently, whose splitting field has odd degree over k, then $f(x)$ is solvable by radicals.

The next proposition gives an example showing that the converse of Galois's theorem is false in prime characteristic.

Lemma 4.55. *If $k = \mathbb{F}_p(t)$, the field of rational functions over $\mathbb{F}_p$, then $f(x) = x^p - x - t$ has no roots in k.*

Proof. If there is a root α of $f(x)$ lying in k, then there are $g(t), h(t) \in \mathbb{F}_p[t]$ with $\alpha = g(t)/h(t)$; we may assume that $(g, h) = 1$. Since α is a root of $f(x)$, we have $(g/h)^p - (g/h) = t$; clearing denominators, there is an equation $g^p - h^{p-1}g = th^p$ in $\mathbb{F}_p[t]$. Hence, $g \mid th^p$. Since $(g, h) = 1$, we have $g \mid t$, so that $g(t) = at$ or $g(t)$ is a constant, say, $g(t) = b$, where $a, b \in \mathbb{F}_p$. Transposing $h^{p-1}g$ in the displayed equation shows that $h \mid g^p$; but $(g, h) = 1$ forces h to be a constant. We conclude that if $\alpha = g/h$, then $\alpha = at$ or $\alpha = b$. In the first case,

$$\begin{aligned} 0 &= \alpha^p - \alpha - t \\ &= (at)^p - (at) - t \\ &= a^p t^p - at - t \\ &= at^p - at - t \qquad \text{by Fermat's theorem in } \mathbb{F}_p \\ &= t(at^{p-1} - a - 1). \end{aligned}$$

It follows that $at^{p-1} - a - 1 = 0$. But $a \neq 0$, and this contradicts t being transcendental over $\mathbb{F}_p$. In the second case, $\alpha = b \in \mathbb{F}_p$. But b is not a root of $f(x)$, for $f(b) = b^p - b - t = -t$, by Fermat's theorem. Thus, no root α of $f(x)$ can lie in k. •

Proposition 4.56. *Let p be a prime, and let $k = \mathbb{F}_p(t)$. The Galois group of $f(x) = x^p - x - t$ over k is cyclic of order p, but $f(x)$ is not solvable by radicals over k.*

Proof. Let α be a root of $f(x)$. It is easy to see that the roots of $f(x)$ are $\alpha + i$, where $0 \leq i < p$, for Fermat's theorem gives $i^p = i$ in $\mathbb{F}_p$, and so

$$(\alpha + i)^p - (\alpha + i) - t = \alpha^p + i^p - \alpha - i - t = \alpha^p - \alpha - t = 0.$$

It follows that $f(x)$ is a separable polynomial and that $k(\alpha)$ is a splitting field of $f(x)$ over k. We claim that $f(x)$ is irreducible in $k[x]$. Suppose that $f(x) = g(x)h(x)$, where

$$g(x) = x^d + c_{d-1}x^{d-1} + \cdots + c_0 \in k[x]$$

and $0 < d < \deg(f) = p$; then $g(x)$ is a product of d factors of the form $\alpha + i$. Now $-c_{d-1} \in k$ is the sum of the roots: $-c_{d-1} = d\alpha + j$, where $j \in \mathbb{F}_p$, and so $d\alpha \in k$. Since $0 < d < p$, however, $d \neq 0$ in k, and this forces $\alpha \in k$, contradicting the lemma. Therefore, $f(x)$ is an irreducible polynomial in $k[x]$. Since $\deg(f) = p$, we have $[k(\alpha) : k] = p$ and, since $f(x)$ is separable, we have $|\operatorname{Gal}(k(\alpha)/k)| = [k(\alpha) : k] = p$. Therefore, $\operatorname{Gal}(k(\alpha)/k) \cong \mathbb{I}_p$.

It will be convenient to have certain roots of unity available. Let Ω be the set of all qth roots of unity, where $q < p$ is a prime divisor of $p!$. We claim that $\alpha \notin k(\Omega)$. On

the one hand, if $n = \prod_{q<p} q$, then Ω is contained in the splitting field of $x^n - 1$, and so $[k(\Omega) : k] \mid n!$, by Theorem 4.3. It follows that $p \nmid [k(\Omega) : k]$. On the other hand, if $\alpha \in k(\Omega)$, then $k(\alpha) \subseteq k(\Omega)$ and $[k(\Omega) : k] = [k(\Omega) : k(\alpha)][k(\alpha) : k] = p[k(\Omega) : k(\alpha)]$. Hence, $p \mid [k(\Omega) : k]$, and this is a contradiction.

If $f(x)$ were solvable by radicals over $k(\Omega)$, there would be a radical extension

$$k(\Omega) = B_0 \subseteq B_1 \subseteq \cdots \subseteq B_r$$

with $k(\Omega, \alpha) \subseteq B_r$. We may assume, for each $i \geq 1$, that B_i/B_{i-1} is of prime type; that is, $B_i = B_{i-1}(u_i)$, where $u_i^{q_i} \in B_{i-1}$ and q_i is prime. There is some $j \geq 1$ with $\alpha \in B_j$ but $\alpha \notin B_{j-1}$. Simplifying notation, we set $u_j = u$, $q_j = q$, $B_{j-1} = B$, and $B_j = B'$. Thus, $B' = B(u)$, $u^q = b \in B$, $\alpha \in B'$, and $\alpha, u \notin B$. We claim that $f(x) = x^p - x - t$, which we know to be irreducible in $k[x]$, is also irreducible in $B[x]$. By accessory irrationalities, Exercise 4.5 on page 217, restriction gives an injection $\text{Gal}(B(\alpha)/B) \to \text{Gal}(k(\alpha)/k)) \cong \mathbb{I}_p$. If $\text{Gal}(B(\alpha)/B) = \{1\}$, then $B(\alpha) = B$ and $\alpha \in B$, a contradiction. Therefore, $\text{Gal}(B(\alpha)/B) \cong \mathbb{I}_p$, and $f(x)$ is irreducible in $B[x]$, by Exercise 4.11 on page 218.

Since $u \notin B'$ and B contains all the qth roots of unity, Proposition 3.126 shows that $x^q - b$ is irreducible in $B[x]$, for it does not split in $B[x]$. Now $B' = B(u)$ is a splitting field of $x^q - b$, and so $[B' : B] = q$. We have $B \subsetneq B(\alpha) \subseteq B'$, and

$$q = [B' : B] = [B' : B(\alpha)][B(\alpha) : B].$$

Since q is prime, $[B' : B(\alpha)] = 1$; that is, $B' = B(\alpha)$, and so $q = [B' : B]$. As α is a root of the irreducible polynomial $f(x) = x^p - x - t \in B[x]$, we have $[B(\alpha) : B] = p$; therefore, $q = p$. Now $B(u) = B' = B(\alpha)$ is a separable extension, by Proposition 4.38, for α is a separable element. It follows that $u \in B'$ is also a separable element, contradicting $\text{irr}(u, B) = x^q - b = x^p - b = (x - u)^p$ having repeated roots.

We have shown that $f(x)$ is not solvable by radicals over $k(\Omega)$. It follows that $f(x)$ is not solvable by radicals over k, for if there were a radical extension $k = R_0 \subseteq R_1 \subseteq \cdots \subseteq R_t$ with $k(\alpha) \subseteq R_t$, then $k(\Omega) = R_0(\Omega) \subseteq R_1(\Omega) \subseteq \cdots \subseteq R_t(\Omega)$ would show that $f(x)$ is solvable by radicals over $k(\Omega)$, a contradiction. •

The *discriminant* of a polynomial is useful in computing its Galois group.

Definition. If $f(x) = \prod_i (x - \alpha_i) \in k[x]$, where k is a field, define

$$\Delta = \prod_{i<j} (\alpha_i - \alpha_j),$$

and define the ***discriminant*** to be $D = D(f) = \Delta^2 = \prod_{i<j} (\alpha_i - \alpha_j)^2$.

It is clear that $f(x)$ has repeated roots if and only if its discriminant $D = 0$.

The product $\Delta = \prod_{i<j} (\alpha_i - \alpha_j)$ has one factor $\alpha_i - \alpha_j$ for each distinct pair of indices (i, j) (the restriction $i < j$ prevents a pair of indices from occurring twice). If E/k is a splitting field of $f(x)$ and if $G = \text{Gal}(E/k)$, then each $\sigma \in G$ permutes the roots, and so

σ permutes all the distinct pairs. However, it may happen that $i < j$ while the subscripts involved in $\sigma(\alpha_i) - \sigma(\alpha_j)$ are in reverse order. For example, suppose the roots of a cubic are α_1, α_2, and α_3, and suppose there is $\sigma \in G$ with $\sigma(\alpha_1) = \alpha_2$, $\sigma(\alpha_2) = \alpha_1$, and $\sigma(\alpha_3) = \alpha_3$. Then

$$\begin{aligned}\sigma(\Delta) &= \big(\sigma(\alpha_1) - \sigma(\alpha_2)\big)\big(\sigma(\alpha_1) - \sigma(\alpha_3)\big)\big(\sigma(\alpha_2) - \sigma(\alpha_3)\big)\\ &= (\alpha_2 - \alpha_1)(\alpha_2 - \alpha_3)(\alpha_1 - \alpha_3)\\ &= -(\alpha_1 - \alpha_2)(\alpha_2 - \alpha_3)(\alpha_1 - \alpha_3)\\ &= -\Delta.\end{aligned}$$

In general, each term $\alpha_i - \alpha_j$ occurs in $\sigma(\Delta)$ with a possible sign change. We conclude, for all $\sigma \in \text{Gal}(E/k)$, that $\sigma(\Delta) = \pm\Delta$. It is natural to consider Δ^2 rather than Δ, for Δ depends not only on the roots of $f(x)$, but also on the order in which they are listed, whereas $D = \Delta^2$ does not depend on the listing of the roots. For a connection between discriminants and the alternating group A_n, see Proposition 4.59(ii) on page 241.

Proposition 4.57. *If $f(x) \in k[x]$ is a separable polynomial, then its discriminant D lies in k.*

Proof. Let E/k be a splitting field of $f(x)$; since $f(x)$ is separable, Theorem 4.34 applies to show that E/k is a Galois extension. Each $\sigma \in \text{Gal}(E/k)$ permutes the roots $u_1, \ldots, u_n$ of $f(x)$, and $\sigma(\Delta) = \pm\Delta$, as we have just seen. Therefore,

$$\sigma(D) = \sigma(\Delta^2) = \sigma(\Delta)^2 = (\pm\Delta)^2 = D,$$

so that $D \in E^G$. Since E/k is a Galois extension, we have $E^G = k$, and so $D \in k$. •

If $f(x) = x^2 + bx + c$, then the quadratic formula gives the roots of $f(x)$:

$$\alpha = \tfrac{1}{2}\big(-b + \sqrt{b^2 - 4c}\big) \quad \text{and} \quad \beta = \tfrac{1}{2}\big(-b - \sqrt{b^2 - 4c}\big).$$

It follows that

$$D = \Delta^2 = (\alpha - \beta)^2 = b^2 - 4c.$$

If $f(x)$ is a cubic with roots α, β, γ, then

$$D = \Delta^2 = (\alpha - \beta)^2(\alpha - \gamma)^2(\beta - \gamma)^2;$$

it is not obvious how to compute the discriminant D from the coefficients of $f(x)$.

Definition. A polynomial $f(x) = x^n + c_{n-1}x^{n-1} + \cdots + c_0 \in k[x]$ is ***reduced*** if $c_{n-1} = 0$. If $f(x)$ is a monic polynomial of degree n and if $c_{n-1} \neq 0$ in k, where $\text{char}(k) = 0$, then its ***associated reduced polynomial*** is

$$\widetilde{f}(x) = f(x - \tfrac{1}{n}c_{n-1}).$$

If $f(x) = x^n + c_{n-1}x^{n-1} + \cdots + c_0 \in k[x]$ and $\beta \in k$ is a root of $\widetilde{f}(x)$, then

$$0 = \widetilde{f}(\beta) = f(\beta - \tfrac{1}{n}c_{n-1}).$$

Hence, β is a root of $\widetilde{f}(x)$ if and only if $\beta - \frac{1}{n}c_{n-1}$ is a root of $f(x)$.

Theorem 4.58. *Let k be a field of characteristic 0.*

(i) *A polynomial $f(x) \in k[x]$ and its associated reduced polynomial $\widetilde{f}(x)$ have the same discriminant.*

(ii) *The discriminant of a reduced cubic $\widetilde{f}(x) = x^3 + qx + r$ is*

$$D = -4q^3 - 27r^2.$$

Proof. (i) If the roots of $f(x) = \sum c_i x^i$ are $\alpha_1, \ldots, \alpha_n$, then the roots of $\widetilde{f}(x)$ are $\beta_1, \ldots, \beta_n$, where $\beta_i = \alpha_i + \frac{1}{n}c_{n-1}$. Therefore, $\beta_i - \beta_j = \alpha_i - \alpha_j$ for all i, j,

$$\prod_{i<j}(\alpha_i - \alpha_j) = \prod_{i<j}(\beta_i - \beta_j),$$

and so the discriminants, which are the squares of these, are equal.

(ii) The cubic formula gives the roots of $\widetilde{f}(x)$ as

$$\alpha = g + h, \quad \beta = \omega g + \omega^2 h, \quad \text{and} \quad \gamma = \omega^2 g + \omega h,$$

where $g = \left[\frac{1}{2}\left(-r + \sqrt{R}\right)\right]^{1/3}$, $h = -q/3g$, $R = r^2 + \frac{4}{27}q^3$, and ω is a cube root of unity. Because $\omega^3 = 1$, we have

$$\begin{aligned} \alpha - \beta &= (g + h) - (\omega g + \omega^2 h) \\ &= (g - \omega^2 h) - (\omega g - h) \\ &= (g - \omega^2 h) - (g - \omega^2 h)\omega \\ &= (g - \omega^2 h)(1 - \omega). \end{aligned}$$

Similar calculations give

$$\alpha - \gamma = (g + h) - (\omega^2 g + \omega h) = (g - \omega h)(1 - \omega^2)$$

and

$$\beta - \gamma = (\omega g + \omega^2 h) - (\omega^2 g + \omega h) = (g - h)\omega(1 - \omega).$$

It follows that

$$\Delta = (g - h)(g - \omega h)(g - \omega^2 h)\omega(1 - \omega^2)(1 - \omega)^2.$$

By Exercise 4.14 on page 246, we have $\omega(1-\omega^2)(1-\omega)^2 = 3i\sqrt{3}$; moreover, the identity

$$x^3 - 1 = (x - 1)(x - \omega)(x - \omega^2),$$

with $x = g/h$, gives

$$(g - h)(g - \omega h)(g - \omega^2 h) = g^3 - h^3 = \sqrt{R}$$

(we saw on page 208 that $g^3 - h^3 = \sqrt{R}$). Therefore, $\Delta = 3i\sqrt{3}\sqrt{R}$, and

$$D = \Delta^2 = -27R = -27r^2 - 4q^3. \quad \bullet$$

Remark. Let k be a field, and let $f(x) = a_m x^m + a_{m-1}x^{m-1} + \cdots + a_1 x + a_0$ and $g(x) = b_n x^n + b_{n-1}x^{n-1} + \cdots + b_1 x + b_0 \in k[x]$. Their ***resultant*** is defined as

$$\mathrm{Res}(f, g) = \det(M),$$

where $M = M(f, g)$ is the $(m+n) \times (m+n)$ matrix

$$M = \begin{bmatrix} a_m & a_{m-1} & \cdots & a_1 & a_0 & & & \\ & a_m & a_{m-1} & \cdots & a_1 & a_0 & & \\ & & a_m & a_{m-1} & \cdots & a_1 & a_0 & \\ & & & \cdots & & & & \\ b_n & b_{n-1} & \cdots & b_1 & b_0 & & & \\ & b_n & b_{n-1} & \cdots & b_1 & b_0 & & \\ & & b_n & b_{n-1} & \cdots & b_1 & b_0 & \\ & & & \cdots & & & & \end{bmatrix};$$

there are n rows for the coefficients a_i of $f(x)$ and m rows for the coefficients b_j of $g(x)$; all the entries other than those shown are assumed to be 0. It can be proved that $\mathrm{Res}(f, g) = 0$ if and only if f and g have a nonconstant common divisor. We mention the resultant here because the discriminant can be computed in terms of it:

$$D(f) = (-1)^{n(n-1)/2}\mathrm{Res}(f, f'),$$

where $f'(x)$ is the derivative of $f(x)$. See the exercises in Dummit and Foote, *Abstract Algebra*, pages 600–602. ◀

Here is a way to use the discriminant in computing Galois groups.

Proposition 4.59. *Let k be a field with characteristic $\neq 2$, let $f(x) \in k[x]$ be a polynomial of degree n with no repeated roots, and let $D = \Delta^2$ be its discriminant. Let E/k be a splitting field of $f(x)$, and let $G = \mathrm{Gal}(E/k)$ be regarded as a subgroup of S_n (as in* Theorem 4.3).

(i) *If $H = A_n \cap G$, then $E^H = k(\Delta)$.*

(ii) *G is a subgroup of A_n if and only if $\sqrt{D} \in k$.*

Proof. (i) The second isomorphism theorem gives $H = (G \cap A_n) \lhd G$ and

$$[G : H] = [G : A_n \cap G] = [A_n G : A_n] \leq [S_n : A_n] = 2.$$

By the fundamental theorem of Galois theory (which applies because $f(x)$ has no repeated roots, hence is separable), $[E^H : k] = [G : H]$, so that $[E^H : k] = [G : H] \leq 2$. By Exercise 4.25 on page 248, we have $k(\Delta) \subseteq E^{A_n}$, and so $k(\Delta) \subseteq E^H$. Therefore,

$$[E^H : k] = [E^H : k(\Delta)][k(\Delta) : k] \leq 2. \tag{5}$$

There are two cases. If $[E^H : k] = 1$, then each factor in Eq. (5) is 1; in particular, $[E^H : k(\Delta)] = 1$ and $E^H = k(\Delta)$. If $[E^H : k] = 2$, then $[G : H] = 2$ and there exists $\sigma \in G$, $\sigma \notin A_n$, so that $\sigma(\Delta) = -\Delta$. Now $\Delta \neq 0$, because $f(x)$ has no repeated roots, and $-\Delta \neq \Delta$, because k does not have characteristic 2. Hence, $\Delta \notin E^G = k$ and $[k(\Delta) : k] > 1$. It follows from Eq. (5) that $[E^H : k(\Delta)] = 1$ and $E^H = k(\Delta)$.

(ii) The following are equivalent: $G \leq A_n$; $H = G \cap A_n = G$; $E^H = E^G = k$. Since $E^H = k(\Delta)$, by part (i), $E^H = k$ is equivalent to $k(\Delta) = k$; that is, $\Delta = \sqrt{D} \in k$. •

We now show how to compute Galois groups of polynomials over $\mathbb{Q}$ of low degree.

If $f(x) \in \mathbb{Q}[x]$ is quadratic, then its Galois group has order either 1 or 2 (because the symmetric group S_2 has order 2). The Galois group has order 1 if $f(x)$ splits; it has order 2 if $f(x)$ does not split; that is, if $f(x)$ is irreducible.

If $f(x) \in \mathbb{Q}[x]$ is a cubic having a rational root, then its Galois group G is the same as that of its quadratic factor. Otherwise $f(x)$ is irreducible; since $|G|$ is now a multiple of 3, by Corollary 4.9, and $G \leq S_3$, it follows that either $G \cong A_3 \cong \mathbb{I}_3$ or $G \cong S_3$.

Proposition 4.60. *Let $f(x) \in \mathbb{Q}[x]$ be an irreducible cubic with Galois group G and discriminant D.*

(i) *$f(x)$ has exactly one real root if and only if $D < 0$, in which case $G \cong S_3$.*

(ii) *$f(x)$ has three real roots if and only if $D > 0$. In this case, either $\sqrt{D} \in \mathbb{Q}$ and $G \cong \mathbb{I}_3$, or $\sqrt{D} \notin \mathbb{Q}$ and $G \cong S_3$.*

Proof. Note first that $D \neq 0$: Since $\mathbb{Q}$ has characteristic 0, irreducible polynomials over $\mathbb{Q}$ have no repeated roots. If $f(x)$ has three real roots, then Δ is real and $D = \Delta^2 > 0$. The other possibility is that $f(x)$ has one real root α and two complex roots: $\beta = u + iv$ and $\overline{\beta} = u - iv$. Since $\beta - \overline{\beta} = 2iv$ and $\alpha = \overline{\alpha}$, we have

$$\begin{aligned}\Delta &= (\alpha - \beta)(\alpha - \overline{\beta})(\beta - \overline{\beta})\\ &= (\alpha - \beta)(\overline{\alpha - \beta})(\beta - \overline{\beta})\\ &= |\alpha - \beta|^2(2iv),\end{aligned}$$

and so $D = \Delta^2 = -4v^2|\alpha - \beta|^4 < 0$.

Let $E/\mathbb{Q}$ be the splitting field of $f(x)$. If $f(x)$ has exactly one real root α, then $E \neq \mathbb{Q}(\alpha)$. Hence $|G| > 3$ and $G \cong S_3$. If $f(x)$ has three real roots, then $D > 0$ and $\sqrt{D}$ is real. By Proposition 4.59(ii), $G \cong A_3 \cong \mathbb{I}_3$ if and only if $\sqrt{D}$ is rational; hence $G \cong S_3$ if $\sqrt{D}$ is irrational. •

Example 4.61.

The polynomial $f(x) = x^3 - 2 \in \mathbb{Q}[x]$ is irreducible, by Theorem 3.43. Its discriminant is $D = -108$, and so it has one real root; since $\sqrt{-108} \notin \mathbb{Q}$ (it is not even real), the Galois group of $f(x)$ is not contained in A_3. Thus, the Galois group is S_3.

The polynomial $x^3 - 4x + 2 \in \mathbb{Q}[x]$ is irreducible, by Theorem 3.43 or by Eisenstein's criterion; its discriminant is $D = 148$, and so it has 3 real roots. Since $\sqrt{148}$ is irrational, the Galois group is S_3.

The polynomial $f(x) = x^3 - 48x + 64 \in \mathbb{Q}[x]$ is irreducible, by Theorem 3.43; the discriminant is $D = 2^{12}3^4$, and so $f(x)$ has 3 real roots. Since $\sqrt{D}$ is rational, the Galois group is $A_3 \cong \mathbb{I}_3$. ◀

Before examining quartics, let us note that if d is a divisor of $|S_4| = 24$, then it is known that S_4 has a subgroup of order d (see Exercise 5.23 on page 277). If $d = 4$, then $\mathbf{V}$ and $\mathbb{I}_4$ are nonisomorphic subgroups of order d; for any other divisor d, any two subgroups of order d are isomorphic. We conclude that the Galois group G of a quartic is determined to isomorphism by its order unless $|G| = 4$.

Consider a (reduced) quartic $f(x) = x^4 + qx^2 + rx + s \in \mathbb{Q}[x]$; let $E/\mathbb{Q}$ be its splitting field and let $G = \mathrm{Gal}(E/\mathbb{Q})$ be its Galois group. [By Exercise 4.15 on page 246, there is no loss in generality in assuming that $f(x)$ is reduced.] If $f(x)$ has a rational root α, then $f(x) = (x - \alpha)c(x)$, and its Galois group is the same as that of the cubic factor $c(x)$; but Galois groups of cubics have already been discussed. Suppose that $f(x) = h(x)\ell(x)$ is the product of two irreducible quadratics; let α be a root of $h(x)$ and let β be a root of $\ell(x)$. If $\mathbb{Q}(\alpha) \cap \mathbb{Q}(\beta) = \mathbb{Q}$, then Exercise 4.17(iv) on page 246 shows that $G \cong \mathbf{V}$, the four group; otherwise, $\alpha \in \mathbb{Q}(\beta)$, so that $\mathbb{Q}(\beta) = \mathbb{Q}(\alpha, \beta) = E$, and G has order 2.

We are left with the case $f(x)$ irreducible. The basic idea now is to compare G with the four group $\mathbf{V}$, namely, the normal subgroup of S_4

$$\mathbf{V} = \big\{(1), (1\ 2)(3\ 4), (1\ 3)(2\ 4), (1\ 4)(2\ 3)\big\},$$

so that we can identify the fixed field of $\mathbf{V} \cap G$. If the four (necessarily distinct) roots of $f(x)$ are $\alpha_1, \alpha_2, \alpha_3, \alpha_4$, consider the numbers [which are distinct, by Proposition 4.63(ii)]:

$$\begin{cases} u = (\alpha_1 + \alpha_2)(\alpha_3 + \alpha_4), \\ v = (\alpha_1 + \alpha_3)(\alpha_2 + \alpha_4), \\ w = (\alpha_1 + \alpha_4)(\alpha_2 + \alpha_3). \end{cases} \tag{6}$$

It is clear that if $\sigma \in \mathbf{V} \cap G$, then σ fixes u, v, and w. Conversely, if $\sigma \in S_4$ fixes $u = (\alpha_1 + \alpha_2)(\alpha_3 + \alpha_4)$, then

$$\sigma \in \mathbf{V} \cup \big\{(1\ 2), (3\ 4), (1\ 3\ 2\ 4), (1\ 4\ 2\ 3)\big\}.$$

However, none of the last four permutations fixes both v and w, and so $\sigma \in G$ fixes each of u, v, w if and only if $\sigma \in \mathbf{V} \cap G$. Therefore,

$$E^{\mathbf{V} \cap G} = \mathbb{Q}(u, v, w).$$

Definition. The ***resolvent cubic*** of $f(x) = x^4 + qx^2 + rx + s$ is

$$g(x) = (x - u)(x - v)(x - w),$$

where u, v, w are the numbers defined in Eqs. (6).

Proposition 4.62. *The resolvent cubic of* $f(x) = x^4 + qx^2 + rx + s$ *is*

$$g(x) = x^3 - 2qx^2 + (q^2 - 4s)x + r^2.$$

Proof. If $f(x) = (x^2 + jx + \ell)(x^2 - jx + m)$, then we saw, in our discussion of the quartic formula on page 209, that j^2 is a root of

$$h(x) = x^3 + 2qx^2 + (q^2 - 4s)x - r^2,$$

a polynomial differing from the claimed expression for $g(x)$ only in the sign of its quadratic and constant terms. Thus, a number β is a root of $h(x)$ if and only if $-\beta$ is a root of $g(x)$.

Let the four roots $\alpha_1, \alpha_2, \alpha_3, \alpha_4$ of $f(x)$ be indexed so that α_1, α_2 are roots of $x^2+jx+\ell$ and α_3, α_4 are roots of x^2-jx+m. Then $j = -(\alpha_1+\alpha_2)$ and $-j = -(\alpha_3+\alpha_4)$; therefore,

$$u = (\alpha_1 + \alpha_2)(\alpha_3 + \alpha_4) = -j^2$$

and $-u$ is a root of $h(x)$ since $h(j^2) = 0$.

Now factor $f(x)$ into two quadratics, say,

$$f(x) = (x^2 + \tilde{j}x + \tilde{\ell})(x^2 - \tilde{j}x + \tilde{m}),$$

where α_1, α_3 are roots of the first factor and α_2, α_4 are roots of the second. The same argument as before now shows that

$$v = (\alpha_1 + \alpha_3)(\alpha_2 + \alpha_4) = -\tilde{j}^2;$$

hence $-v$ is a root of $h(x)$. Similarly, $-w = -(\alpha_1 + \alpha_4)(\alpha_2 + \alpha_3)$ is a root of $h(x)$. Therefore,

$$h(x) = (x + u)(x + v)(x + w),$$

and so

$$g(x) = (x - u)(x - v)(x - w)$$

is obtained from $h(x)$ by changing the sign of the quadratic and constant terms. •

Proposition 4.63.

(i) *The discriminant* $D(f)$ *of a quartic polynomial* $f(x) \in \mathbb{Q}[x]$ *is equal to the discriminant* $D(g)$ *of its resolvent cubic* $g(x)$.

(ii) *If* $f(x)$ *is irreducible, then* $g(x)$ *has no repeated roots.*

Proof. (i) One checks easily that

$$u - v = \alpha_1\alpha_3 + \alpha_2\alpha_4 - \alpha_1\alpha_2 - \alpha_3\alpha_4 = -(\alpha_1 - \alpha_4)(\alpha_2 - \alpha_3).$$

Similarly,

$$u - w = -(\alpha_1 - \alpha_3)(\alpha_2 - \alpha_4) \quad \text{and} \quad v - w = (\alpha_1 - \alpha_2)(\alpha_3 - \alpha_4).$$

We conclude that $D(g) = [(u - v)(u - w)(v - w)]^2 = \big[-\prod_{i<j}(\alpha_i - \alpha_j)\big]^2 = D(f)$.

(ii) If $f(x)$ is irreducible, then it has no repeated roots (for it is separable because $\mathbb{Q}$ has characteristic 0), and so $D(f) \neq 0$. Therefore, $D(g) = D(f) \neq 0$, and so $g(x)$ has no repeated roots. •

In the notation of Eqs. (6), if $f(x)$ is an irreducible quartic, then u, v, w are distinct.

Proposition 4.64. *Let $f(x) \in \mathbb{Q}[x]$ be an irreducible quartic with Galois group G with discriminant D, and let m be the order of the Galois group of its resolvent cubic $g(x)$.*

(i) *If $m = 6$, then $G \cong S_4$. In this case, $g(x)$ is irreducible and $\sqrt{D}$ is irrational.*

(ii) *If $m = 3$, then $G \cong A_4$. In this case, $g(x)$ is irreducible and $\sqrt{D}$ is rational.*

(iii) *If $m = 1$, then $G \cong \mathbf{V}$. In this case, $g(x)$ splits in $\mathbb{Q}[x]$.*

(iv) *If $m = 2$, then $G \cong D_8$ or $G \cong \mathbb{I}_4$. In this case, $g(x)$ has an irreducible quadratic factor.*

Proof. We have seen that $E^{\mathbf{V}\cap G} = \mathbb{Q}(u, v, w)$. By the fundamental theorem of Galois theory,

$$\begin{aligned}[G : \mathbf{V} \cap G] &= [E^{\mathbf{V}\cap G} : \mathbb{Q}] \\ &= [\mathbb{Q}(u, v, w) : \mathbb{Q}] \\ &= |\operatorname{Gal}(\mathbb{Q}(u, v, w)/\mathbb{Q})| \\ &= m.\end{aligned}$$

Since $f(x)$ is irreducible, $|G|$ is divisible by 4, by Corollary 4.9, and the group-theoretic statements follow from Exercise 4.28 on page 248 and Exercise 4.29 on page 248. Finally, in the first two cases, $|G|$ is divisible by 12, and Proposition 4.59(ii) decides whether $G \cong S_4$ or $G \cong A_4$. The conditions on $g(x)$ in the last two last two cases are easy to see. •

We have seen that the resolvent cubic has much to say about the Galois group of the irreducible quartic from which it comes.

Example 4.65.

(i) Let $f(x) = x^4 - 4x + 2 \in \mathbb{Q}[x]$; $f(x)$ is irreducible [the best way to see this is with Eisenstein's criterion, Theorem 6.34, but we can also see that $f(x)$ has no rational roots, using Theorem 3.43, and then showing that $f(x)$ has no irreducible quadratic factors by examining conditions imposed on its coefficients]. By Proposition 4.62, the resolvent cubic is

$$g(x) = x^3 - 8x + 16.$$

Now $g(x)$ is irreducible (again, the best way to see this uses some results of Chapter 6: specifically, Theorem 6.30, for if we reduce mod 5, we obtain $x^3 + 2x + 1$, and this polynomial is irreducible over $\mathbb{I}_5$ because it has no roots). The discriminant of $g(x)$ is -4864, so that Theorem 4.60 shows that the Galois group of $g(x)$ is S_3, hence has order 6. Theorem 4.64 now shows that $G \cong S_4$.

(ii) Let $f(x) = x^4 - 10x^2 + 1 \in \mathbb{Q}[x]$; $f(x)$ is irreducible, by Exercise 6.23(viii) on page 339. By Proposition 4.62, the resolvent cubic is

$$x^3 + 20x^2 + 96x = x(x + 8)(x + 12).$$

In this case, $\mathbb{Q}(u, v, w) = \mathbb{Q}$ and $m = 1$. Therefore, $G \cong \mathbf{V}$. [This should not be a surprise if we recall Example 3.122, where we saw that $f(x)$ arises as the irreducible polynomial of $\alpha = \sqrt{2} + \sqrt{3}$, where $\mathbb{Q}(\alpha) = \mathbb{Q}(\sqrt{2}, \sqrt{3})$.] ◂

An interesting open question is the ***inverse Galois problem***: Which finite abstract groups G are isomorphic to $\mathrm{Gal}(E/\mathbb{Q})$, where $E/\mathbb{Q}$ is a Galois extension? D. Hilbert proved that the symmetric groups S_n are such Galois groups, and I. Shafarevich proved that every solvable group is a Galois group (see Neukirch–Schmidt–Wingberg, *Cohomology of Number Fields*). After the classification of the finite simple groups in the 1980s, it was shown that most simple groups are Galois groups. For more information, the reader is referred to Malle–Matzat, *Inverse Galois Theory*.

EXERCISES

4.14 Prove that $\omega(1-\omega^2)(1-\omega)^2 = 3i\sqrt{3}$, where $\omega = e^{2\pi i/3}$.

4.15 (i) Prove that if $a \neq 0$, then $f(x)$ and $af(x)$ have the same discriminant and the same Galois group. Conclude that it is no loss in generality to restrict attention to monic polynomials when computing Galois groups.

(ii) Let k be a field of characteristic 0. Prove that a polynomial $f(x) \in k[x]$ and its associated reduced polynomial $\widetilde{f}(x)$ have the same Galois group.

4.16 (i) Let k be a field of characteristic 0. If $f(x) = x^3 + ax^2 + bx + c \in k[x]$, then its associated reduced polynomial is $x^3 + qx + r$, where

$$q = b - \tfrac{1}{3}a^2 \quad \text{and} \quad r = \tfrac{2}{27}a^3 - \tfrac{1}{3}ab + c.$$

(ii) Show that the discriminant of $f(x)$ is

$$D = a^2b^2 - 4b^3 - 4a^3c - 27c^2 + 18abc.$$

4.17 Let k be a field, let $f(x) \in k[x]$ be a separable polynomial, and let E/k be a splitting field of $f(x)$. Assume further that there is a factorization

$$f(x) = g(x)h(x)$$

in $k[x]$, and that B/k and C/k are intermediate fields that are splitting fields of $g(x)$ and $h(x)$, respectively.

(i) Prove that $\mathrm{Gal}(E/B)$ and $\mathrm{Gal}(E/C)$ are normal subgroups of $\mathrm{Gal}(E/k)$.

(ii) Prove that $\mathrm{Gal}(E/B) \cap \mathrm{Gal}(E/C) = \{1\}$.

(iii) If $B \cap C = k$, prove that $\mathrm{Gal}(E/B)\,\mathrm{Gal}(E/C) = \mathrm{Gal}(E/k)$. (Intermediate fields B and C are called ***linearly disjoint*** if $B \cap C = k$.)

(iv) Use Proposition 2.80 and Theorem 4.16 to show, in this case, that

$$\mathrm{Gal}(E/k) \cong \mathrm{Gal}(B/k) \times \mathrm{Gal}(C/k).$$

(Note that $\mathrm{Gal}(B/k)$ is not a subgroup of $\mathrm{Gal}(E/k)$.)

(v) Use (iv) to give another proof that $\mathrm{Gal}(E/\mathbb{Q}) \cong \mathbf{V}$, where $E = \mathbb{Q}(\sqrt{2} + \sqrt{3})$ [see Example 3.122 on page 190].

(vi) Let $f(x) = (x^3 - 2)(x^3 - 3) \in \mathbb{Q}[x]$. If $B/\mathbb{Q}$ and $C/\mathbb{Q}$ are the splitting fields of $x^3 - 2$ and $x^3 - 3$ inside $\mathbb{C}$, prove that $\mathrm{Gal}(E/\mathbb{Q}) \not\cong \mathrm{Gal}(B/\mathbb{Q}) \times \mathrm{Gal}(C/\mathbb{Q})$, where E is the splitting field of $f(x)$ contained in $\mathbb{C}$.

4.18 Let k be a field of characteristic 0, and let $f(x) \in k[x]$ be a polynomial of degree 5 with splitting field E/k. Prove that $f(x)$ is solvable by radicals if and only if $[E : k] < 60$.

4.19 (i) If $\mathcal{L}$ and $\mathcal{L}'$ are lattices, a function $f : \mathcal{L} \to \mathcal{L}'$ is called ***order-preserving*** if $a \preceq b$ in $\mathcal{L}$ implies $f(a) \preceq f(b)$ in $\mathcal{L}'$. Prove that if $\mathcal{L}$ and $\mathcal{L}'$ are lattices and $\varphi : \mathcal{L} \to \mathcal{L}'$ is a bijection such that both φ and φ^{-1} are order-preserving, then

$$\varphi(a \wedge b) = \varphi(a) \wedge \varphi(b) \quad \text{and} \quad \varphi(a \vee b) = \varphi(a) \vee \varphi(b).$$

Hint. Adapt the proof of Lemma 4.42.

(ii) Let E/k be a Galois extension with $\mathrm{Gal}(E/k)$ cyclic of order n. Prove that

$$\varphi : \mathrm{Int}(E/k) \to \mathrm{Div}(n),$$

[see Example 4.40(iv)] defined by $\varphi(L) = [L : k]$, is an order-preserving lattice isomorphism.

(iii) Prove that if L and K are subfields of $\mathbb{F}_{p^n}$, then

$$[L \vee K : \mathbb{F}_p] = \mathrm{lcm}\big([L : \mathbb{F}_p], [K : \mathbb{F}_p]\big)$$

and

$$[L \cap K : \mathbb{F}_p] = \gcd\big([L : \mathbb{F}_p], [K : \mathbb{F}_p]\big).$$

4.20 Find all finite fields k whose subfields form a *chain*; that is, if k' and k'' are subfields of k, then either $k' \subseteq k''$ or $k'' \subseteq k'$.

4.21 (i) Let k be an infinite field, let $f(x) \in k[x]$ be a separable polynomial, and let $E = k(\alpha_1, \ldots, \alpha_n)$, where $\alpha_1, \ldots, \alpha_n$ are the roots of $f(x)$. Prove that there are $c_i \in k$ so that $E = k(\beta)$, where $\beta = c_1\alpha_1 + \cdots + c_n\alpha_n$.

Hint. Use the proof of Steinitz's theorem.

(ii) (***Janusz***). Let k be a finite field and let $E = k(\alpha, \beta)$. Prove that if $k(\alpha)$ and $k(\beta)$ are linearly disjoint [that is, if $k(\alpha) \cap k(\beta) = k$], then $E = k(\alpha + \beta)$. (This result is false in general. For example, N. Boston used the computer algebra system MAGMA to show that there is a primitive element α of $\mathbb{F}_{2^6}$ and a primitive element β of $\mathbb{F}_{2^{10}}$ such that $\mathbb{F}_2(\alpha, \beta) = \mathbb{F}_{2^{30}}$ while $\mathbb{F}_2(\alpha + \beta) = \mathbb{F}_{2^{15}}$.)

Hint. Use Exercise 4.19(iii) and Exercise 1.26 on page 13.

4.22 Let E/k be a Galois extension with Galois group $G = \mathrm{Gal}(E/k)$. Define the ***trace*** $T : E \to E$ by

$$T(u) = \sum_{\sigma \in G} \sigma(u).$$

(i) Prove that $\mathrm{im}\, T \subseteq k$ and that $T(u + v) = T(u) + T(v)$ for all $u, v \in E$.

(ii) Use independence of characters to prove that T is not identically zero.

4.23 Let E/k be a Galois extension with $[E : k] = n$ and with cyclic Galois group $G = \mathrm{Gal}(E/k)$, say, $G = \langle\sigma\rangle$.

(i) Define $\tau = \sigma - 1_E$, and prove that $\ker T = \ker \tau$.

Hint. Show that $\ker \tau = k$, so that $\dim(\mathrm{im}\, \tau) = n - 1 = \dim(\ker T)$.

(ii) ***Trace Theorem***: Prove that if E/k is a Galois extension with cyclic Galois group $\text{Gal}(E/k) = \langle\sigma\rangle$, then

$$\ker T = \{a \in E : a = \sigma(u) - u \text{ for some } u \in E\}.$$

4.24 Let k be a field of characteristic $p > 0$, and let E/k be a Galois extension having a cyclic Galois group $G = \langle\sigma\rangle$ of order p. Using the trace theorem, prove that there is an element $u \in E$ with $\sigma(u) - u = 1$. Prove that $E = k(u)$ and that there is $c \in k$ with $\text{irr}(u, k) = x^p - x - c$.

4.25 If $\sigma \in S_n$ and $f(x_1, \ldots, x_n) \in k[x_1, \ldots, x_n]$, where k is a field, define

$$(\sigma f)(x_1, \ldots, x_n) = f(x_{\sigma 1}, \ldots, x_{\sigma n}).$$

(i) Prove that $(\sigma, f(x_1, \ldots, x_n) \mapsto \sigma f$ defines an action of S_n on $k[x_1, \ldots, x_n]$.

(ii) Let $\Delta = \Delta(x_1, \ldots, x_n) = \prod_{i<j}(x_i - x_j)$ (on page 239, we saw that $\sigma\Delta = \pm\Delta$ for all $\sigma \in S_n$). If $\sigma \in S_n$, prove that $\sigma \in A_n$ if and only if $\sigma\Delta = \Delta$.

Hint. Define $\varphi\colon S_n \to G$, where G is the multiplicative group $\{1, -1\}$, by

$$\varphi(\sigma) = \begin{cases} 1 & \text{if } \sigma\Delta = \Delta; \\ -1 & \text{if } \sigma\Delta = -\Delta. \end{cases}$$

Prove that φ is a homomorphism, and that $\ker\varphi = A_n$.

4.26 Prove that if $f(x) \in \mathbb{Q}[x]$ is an irreducible quartic whose discriminant is rational, then its Galois group has order 4 or 12.

4.27 Let $f(x) = x^4 + rx + s \in \mathbb{Q}[x]$ have Galois group G.

(i) Prove that the discriminant of $f(x)$ is $-27r^4 + 256s^3$.

(ii) Prove that if $s < 0$, then G is not isomorphic to a subgroup of A_4.

(iii) Prove that $f(x) = x^4 + x + 1$ is irreducible and that $G \cong S_4$.

4.28 Let G be a subgroup of S_4 with $|G|$ a multiple of 4, and define $m = |G/(G \cap \mathbf{V})|$.

(i) Prove that m is a divisor of 6.

(ii) If $m = 6$, then $G = S_4$; if $m = 3$, then $G = A_4$; if $m = 1$, then $G = \mathbf{V}$; if $m = 2$, then $G \cong D_8$, $G \cong \mathbb{I}_4$, or $G \cong \mathbf{V}$.

4.29 Let G be a subgroup of S_4. If G acts transitively on $X = \{1, 2, 3, 4\}$ and $|G/(\mathbf{V} \cap G)| = 2$, then $G \cong D_8$ or $G \cong \mathbb{I}_4$. [If we merely assume that G acts transitively on X, then $|G|$ is a multiple of 4 (Corollary 4.9). The added hypothesis $|G/(\mathbf{V} \cap G)| = 2$ removes the possibility $G \cong \mathbf{V}$ when $m = 2$ in Exercise 4.28.]

4.30 Compute the Galois group over $\mathbb{Q}$ of $x^4 + x^2 - 6$.

4.31 Compute the Galois group over $\mathbb{Q}$ of $f(x) = x^4 + x^2 + x + 1$.
Hint. Use Example 3.35(ii) to prove irreducility of $f(x)$, and prove irreducibility of the resolvent cubic by reducing mod 2.

4.32 Compute the Galois group over $\mathbb{Q}$ of $f(x) = 4x^4 + 12x + 9$.
Hint. Prove that $f(x)$ is irreducible in two steps: First show that it has no rational roots, and then use Descartes's method (on page 209) to show that $f(x)$ is not the product of two quadratics over $\mathbb{Q}$.

5

Groups II

We now seek some structural information about groups. Finite abelian groups turn out to be rather uncomplicated: They are direct sums of cyclic groups. Returning to nonabelian groups, the Sylow theorems show, for any prime p, that finite groups G have subgroups of order p^e, where p^e is the largest power of p dividing $|G|$, and any two such are isomorphic. The ideas of normal series and solvability that arose in Galois theory yield invariants of groups (the Jordan–Hölder theorem), showing that simple groups are, in a certain sense, building blocks of finite groups. Consequently, we display more examples of simple groups to accompany the alternating groups A_n, for $n \geq 5$, which we have already proved to be simple. This chapter concludes by investigating free groups and presentations, for they are useful in constructing and describing arbitrary groups. The chapter ends with a proof that every subgroup of a free group is itself a free group.

5.1 Finite Abelian Groups

We continue our study of groups by classifying all finite abelian groups; as is customary, we use the additive notation for the binary operation in these groups. We are going to prove that every finite abelian group is a direct sum of cyclic groups and that this decomposition is unique in a strong sense.

Direct Sums

Groups in this subsection are arbitrary, possibly infinite, abelian groups.

Let us say at the outset that there are two ways to describe the *direct sum* of abelian groups $S_1, \ldots, S_n$. The easiest version is sometimes called their ***external direct sum***, which we denote by $S_1 \times \cdots \times S_n$; its elements are the n-tuples $(s_1, \ldots, s_n)$, where $s_i \in S_i$ for all i, and its binary operation is

$$(s_1, \ldots, s_n) + (s_1', \ldots, s_n') = (s_1 + s_1', \ldots, s_n + s_n').$$

However, the most useful version, isomorphic to $S_1 \times \cdots \times S_n$, is sometimes called their ***internal direct sum***; it involves subgroups S_i of a given group G with $G \cong S_1 \times \cdots \times S_n$. We will usually omit the adjectives *external* and *internal*.

The definition of the direct sum of two subgroups is the additive version of the statement of Proposition 2.80.

Definition. If S and T are subgroups of an abelian group G, then G is the ***direct sum***, denoted by

$$G = S \oplus T,$$

if $S + T = G$ (i.e., for each $a \in G$, there are $s \in S$ and $t \in T$ with $a = s + t$) and $S \cap T = \{0\}$.

Here are several characterizations of a direct sum.

Proposition 5.1. *The following statements are equivalent for an abelian group G and subgroups S and T of G.*

(i) $G = S \oplus T$.

(ii) *Every $g \in G$ has a unique expression of the form*

$$g = s + t,$$

where $s \in S$ and $t \in T$.

(iii) *There are homomorphisms $p\colon G \to S$ and $q\colon G \to T$, called **projections**, and $i\colon S \to G$ and $j\colon T \to G$, called **injections**, such that*

$$pi = 1_S, \quad qj = 1_T, \quad pj = 0, \quad qi = 0, \quad \textit{and} \quad ip + jq = 1_G.$$

Remark. The equations $pi = 1_S$ and $qj = 1_T$ imply that the maps i and j must be injections and the maps p and q must be surjections. ◂

Proof. (i) ⇒ (ii) By hypothesis, $G = S + T$, so that each $g \in G$ has an expression of the form $g = s + t$ with $s \in S$ and $t \in T$. To see that this expression is unique, suppose also that $g = s' + t'$, where $s' \in S$ and $t' \in T$. Then $s + t = s' + t'$ gives $s - s' = t' - t \in S \cap T = \{0\}$. Therefore, $s = s'$ and $t = t'$, as desired.

(ii) ⇒ (iii) If $g \in G$, then there are unique $s \in S$ and $t \in T$ with $g = s + t$. The functions p and q, given by

$$p(g) = s \text{ and } q(g) = t,$$

are well-defined because of the uniqueness hypothesis. It is routine to check that p and q are homomorphisms and that all the equations in the statement hold.

(iii) ⇒ (i) If $g \in G$, the equation $1_G = ip + jq$ gives

$$g = ip(g) + jq(g) \in S + T,$$

because $S = \operatorname{im} i$ and $T = \operatorname{im} j$.

If $g \in S$, then $g = ig$ and $pg = pig = g$; if $g \in T$, then $g = jg$ and $pg = pjg = 0$. Therefore, if $g \in S \cap T$, then $g = 0$. Hence, $S \cap T = \{0\}$, $S + T = G$, and $G = S \oplus T$. •

The next result shows that there is no essential difference between internal and external direct sums.

Corollary 5.2. *Let S and T be subgroups of an abelian group G. If $G = S \oplus T$, then $S \oplus T \cong S \times T$.*

Conversely, given abelian groups S and T, define subgroups $S' \cong S$ and $T' \cong T$ of $S \times T$ by

$$S' = \{(s, 0) : s \in S\} \quad \textit{and} \quad T' = \{(0, t) : t \in T\};$$

then $S \times T = S' \oplus T'$.

Proof. Define $f : S \oplus T \to S \times T$ as follows. If $a \in S \oplus T$, then the proposition says that there is a unique expression of the form $a = s + t$, and so $f : a \mapsto (s, t)$ is a well-defined function. It is routine to check that f is an isomorphism.

Conversely, if $g = (s, t) \in S \times T$, then $g = (s, 0) + (0, t) \in S' + T'$ and $S' \cap T' = \{(0, 0)\}$. Hence, $S \times T = S' \oplus T'$. •

Definition. If $S_1, S_2, \ldots, S_n, \ldots$ are subgroups of an abelian group G, define the ***finite direct sum*** $S_1 \oplus S_2 \oplus \cdots \oplus S_n$ using induction on $n \geq 2$:

$$S_1 \oplus S_2 \oplus \cdots \oplus S_{n+1} = \big[S_1 \oplus S_2 \oplus \cdots \oplus S_n\big] \oplus S_{n+1}.$$

We will also denote the direct sum by

$$\sum_{i=1}^{n} S_i = S_1 \oplus S_2 \oplus \cdots \oplus S_n.$$

Given $S_1, S_2, \ldots, S_n$ subgroups of an abelian group G, when is the subgroup they generate, $\langle S_1, S_2, \ldots, S_n \rangle$, equal to their direct sum? A common mistake is to say that it is enough to assume that $S_i \cap S_j = \{0\}$ for all $i \neq j$, but the following example shows that this is not enough.

Example 5.3.
Let V be a two-dimensional vector space over a field k, which we view as an additive abelian group, and let x, y be a basis. It is easy to check that the intersection of any two of the subspaces $\langle x \rangle$, $\langle y \rangle$, and $\langle x + y \rangle$ is $\{0\}$. On the other hand, we do not have $V = [\langle x \rangle \oplus \langle y \rangle] \oplus \langle x + y \rangle$ because $[\langle x \rangle \oplus \langle y \rangle] \cap \langle x + y \rangle \neq \{0\}$. ◄

In the context of abelian groups, we shall write $S \subseteq G$ to denote S being a subgroup of G, as we do when denoting subrings and ideals; in the context of general, possibly nonabelian, groups, we will continue to write $S \leq G$ to denote a subgroup.

Proposition 5.4. *Let $G = S_1 + S_2 + \cdots + S_n$, where the S_i are subgroups; that is, for each $a \in G$, there are $s_i \in S_i$ for all i, with*

$$a = s_1 + s_2 + \cdots + s_n.$$

Then the following conditions are equivalent.

(i) $G = S_1 \oplus S_2 \oplus \cdots \oplus S_n$.

(ii) *Every $a \in G$ has a unique expression of the form $a = s_1 + s_2 + \cdots + s_n$, where $s_i \in S_i$ for all i.*

(iii) *For each i,*

$$S_i \cap (S_1 + S_2 + \cdots + \widehat{S_i} + \cdots + S_n) = \{0\},$$

where $\widehat{S_i}$ means that the term S_i is omitted from the sum.

Proof. (i) $\Rightarrow$ (ii) The proof is by induction on $n \geq 2$. The base step is Proposition 5.1. For the inductive step, define $T = S_1 + S_2 + \cdots + S_n$, so that $G = T \oplus S_{n+1}$. If $a \in G$, then a has a unique expression of the form $a = t + s_{n+1}$, where $t \in T$ and $s_{n+1} \in S_{n+1}$ (by the proposition). But the inductive hypothesis says that t has a unique expression of the form $t = s_1 + \cdots + s_n$, where $s_i \in S_i$ for all $i \leq n$, as desired.

(ii) $\Rightarrow$ (iii) Suppose that

$$x \in S_i \cap \big(S_1 + S_2 + \cdots + \widehat{S_i} + \cdots + S_n\big).$$

Then $x = s_i \in S_i$ and $s_i = \sum_{j \neq i} s_j$, where $s_j \in S_j$. Unless all the $s_j = 0$, the element 0 has two distinct expressions: $0 = -s_i + \sum_{j \neq i} s_j$ and $0 = 0 + 0 + \cdots + 0$. Therefore, all $s_j = 0$ and $x = s_i = 0$.

(iii) $\Rightarrow$ (i) Since $S_{n+1} \cap \big(S_1 + S_2 + \cdots + S_n\big) = \{0\}$, we have

$$G = S_{n+1} \oplus \big(S_1 + S_2 + \cdots + S_n\big).$$

The inductive hypothesis gives $S_1 + S_2 + \cdots + S_n = S_1 \oplus S_2 \oplus \cdots \oplus S_n$, because, for all $j \leq n$, we have

$$\begin{aligned} S_j \cap \big(S_1 + \cdots + \widehat{S}_j + \cdots + S_n\big) &\subseteq S_j \cap \big(S_1 + \cdots + \widehat{S}_j + \cdots + S_n + S_{n+1}\big) \\ &= \{0\}. \quad \bullet \end{aligned}$$

Corollary 5.5. *Let $G = \langle y_1, \ldots, y_n \rangle$. If, for all $m_i \in \mathbb{Z}$, we have $\sum_i m_i y_i = 0$ implies $m_i y_i = 0$; then*

$$G = \langle y_1 \rangle \oplus \cdots \oplus \langle y_n \rangle.$$

Proof. By Proposition 5.4(ii), it suffices to prove that if $\sum_i k_i y_i = \sum_i \ell_i y_i$, then $k_i y_i = \ell_i y_i$ for all i. But this is clear, for $\sum_i (k_i - \ell_i) y_i = 0$ implies $(k_i - \ell_i) y_i = 0$ for all i. $\bullet$

Example 5.6.
Let V be an n-dimensional vector space over a field k, which we view as an additive abelian group. If $v_1, \ldots, v_n$ is a basis, then

$$V = \langle v_1 \rangle \oplus \langle v_2 \rangle \oplus \cdots \oplus \langle v_n \rangle,$$

where $\langle v_i \rangle = \{rv_i : r \in k\}$ is the one-dimensional subspace spanned by v_i. Each $v \in V$ has a unique expression of the form $v = s_1 + \cdots + s_n$, where $s_i = r_i v_i \in \langle v_i \rangle$, because $v_1, \ldots, v_n$ is a basis. ◂

Now that we have examined finite direct sums, we can generalize Proposition 2.79 from two summands to a finite number of summands. Although we state the result for abelian groups, it should be clear that the proof works for nonabelian groups as well if we assume that the subgroups H_i are normal subgroups (see Exercise 5.1 on page 267).

Proposition 5.7. *If $G_1, G_2, \ldots, G_n$ are abelian groups and $H_i \subseteq G_i$ are subgroups, then*

$$(G_1 \oplus \cdots \oplus G_n)/(H_1 \oplus \cdots \oplus H_n) \cong (G_1/H_1) \times \cdots \times (G_n/H_n).$$

Proof. Define $f : G_1 \oplus \cdots \oplus G_n \to (G_1/H_1) \oplus \cdots \oplus (G_n/H_n)$ by

$$(g_1, \ldots, g_n) \mapsto (g_1 + H_1, \ldots, g_n + H_n).$$

Since f is a surjective homomorphism with $\ker f = H_1 \oplus \cdots \oplus H_n$, the first isomorphism theorem gives the result. •

If G is an abelian group and m is an integer, let us write

$$mG = \{ma : a \in G\}.$$

It is easy to see that mG is a subgroup of G.

Proposition 5.8. *If G is an abelian group and p is a prime, then G/pG is a vector space over $\mathbb{F}_p$.*

Proof. If $[r] \in \mathbb{F}_p$ and $a \in G$, define scalar multiplication

$$[r](a + pG) = ra + pG.$$

This formula is well-defined, for if $k \equiv r \bmod p$, then $k = r + pm$ for some integer m, and so

$$ka + pG = ra + pma + pG = ra + pG,$$

because $pma \in pG$. It is now routine to check that the axioms for a vector space do hold. •

Direct sums of copies of $\mathbb{Z}$ arise often enough to have their own name.

Definition. Let $F = \langle x_1, \ldots, x_n \rangle$ be an abelian group. If

$$F = \langle x_1 \rangle \oplus \cdots \oplus \langle x_n \rangle,$$

where each $\langle x_i \rangle \cong \mathbb{Z}$, then F is called a (finitely generated) ***free abelian group*** with ***basis*** $x_1, \ldots, x_n$. More generally, any group isomorphic to F is called a free abelian group.

For example, $\mathbb{Z}^m = \mathbb{Z} \times \cdots \times \mathbb{Z}$, the group of all m-tuples $(n_1, \ldots, n_m)$ of integers, is a free abelian group.

Proposition 5.9. *If $\mathbb{Z}^m$ denotes the direct sum of m copies of $\mathbb{Z}$, then $\mathbb{Z}^m \cong \mathbb{Z}^n$ if and only if $m = n$.*

Proof. Only necessity needs proof. Note first, for any abelian group G, that if $G = G_1 \oplus \cdots \oplus G_n$, then $2G = 2G_1 \oplus \cdots \oplus 2G_n$. It follows from Proposition 5.7 that

$$G/2G \cong (G_1/2G_1) \oplus \cdots \oplus (G_n/2G_n),$$

so that $|G/2G| = 2^n$. Similarly, if $H = \mathbb{Z}^m$, then $|H/2H| = 2^m$. Finally, if $G = \mathbb{Z}^n \cong \mathbb{Z}^m = H$, then $G/2G \cong H/2H$ and $2^n = 2^m$. We conclude that $n = m$. •

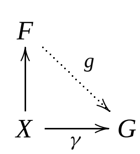
Corollary 5.10. *If F is a (finitely generated) free abelian group, then any two bases of F have the same number of elements.*

Proof. If $x_1, \ldots, x_n$ is a basis of F, then $F \cong \mathbb{Z}^n$, and if $y_1, \ldots, y_m$ is another basis of F, then $F \cong \mathbb{Z}^m$. By the proposition, $m = n$. •

Definition. If F is a free abelian group with basis $x_1, \ldots, x_n$, then n is called the ***rank*** of F, and we write $\operatorname{rank}(F) = n$.

Corollary 5.10 says that $\operatorname{rank}(F)$ is well-defined; that is, it does not depend on the choice of basis. In this language, Proposition 5.9 says that two finitely generated free abelian groups are isomorphic if and only if they have the same rank; that is, the rank of a free abelian group plays the same role as the dimension of a vector space. Comparing the next theorem with Theorem 3.92 shows that a basis of a free abelian group behaves as does a basis of a vector space.

Theorem 5.11. *Let F be a free abelian group with basis $X = \{x_1, \ldots, x_n\}$. If G is any abelian group and if $\gamma \colon X \to G$ is any function, then there exists a unique homomorphism $g \colon F \to G$ with $g(x_i) = \gamma(x_i)$ for all x_i.*



Proof. Every element $a \in F$ has a unique expression of the form

$$a = \sum_{i=1}^{n} m_i x_i,$$

where $m_i \in \mathbb{Z}$. Define $g \colon F \to G$ by

$$g(a) = \sum_{i=1}^{n} m_i \gamma(x_i).$$

If $h \colon F \to G$ is a homomorphism with $h(x_i) = g(x_i)$ for all i, then $h = g$, for two homomorphisms that agree on a set of generators must be equal. •

Theorem 5.11 characterizes free abelian groups.

Proposition 5.12. *Let A be an abelian group containing a subset $X = \{x_1, \ldots, x_n\}$, and let A have the property in* Theorem 5.11*: For every abelian group G and every function $\gamma \colon X \to G$, there exists a unique homomorphism $g \colon A \to G$ with $g(x_i) = \gamma(x_i)$ for all x_i. Then $A \cong \mathbb{Z}^n$; that is, A is a free abelian group of rank n.*

Proof. Consider the diagrams

$$\begin{array}{ccc} A & & \\ {\scriptstyle p}\uparrow & \overset{g}{\searrow} & \\ X & \xrightarrow[q]{} & \mathbb{Z}^n \end{array} \qquad \text{and} \qquad \begin{array}{ccc} \mathbb{Z}^n & & \\ {\scriptstyle q}\uparrow & \overset{h}{\searrow} & \\ X & \xrightarrow[p]{} & A, \end{array}$$

where $p \colon X \to A$ and $q \colon X \to \mathbb{Z}^n$ are inclusions. The first diagram arises from the given property of A, and so $gp = q$; the second arises from Theorem 5.11, which shows that $\mathbb{Z}^n$ enjoys the same property; hence, $hq = p$. We claim that the composite $g \colon A \to \mathbb{Z}^n$ is an isomorphism. To see this, consider the diagram

$$\begin{array}{ccc} A & & \\ {\scriptstyle p}\uparrow & \overset{hg}{\searrow} & \\ X & \xrightarrow[p]{} & A. \end{array}$$

Now $hgp = hq = p$. By hypothesis, hg is the unique such homomorphism. But 1_A is another such, and so $hg = 1_A$. A similar diagram shows that the other composite $gh = 1_{\mathbb{Z}^n}$, and so g is an isomorphism. •

Basis Theorem

It will be convenient to analyze finite abelian groups "one prime at a time."

Recall that a *p-group* is a finite group G of order p^k for some $k \geq 0$. When working wholly in the context of abelian groups, p-groups are called p-primary groups.

Definition. If p is a prime, then an abelian group G is ***p-primary*** if, for each $a \in G$, there is $n \geq 1$ with $p^n a = 0$.

If G is any abelian group, then its ***p-primary component*** is

$$G_p = \{a \in G : p^n a = 0 \text{ for some } n \geq 1\}.$$

It is easy to see, for every prime p, that G_p is a subgroup of G (this is not the case when G is not abelian; for example, G_2 is not a subgroup if $G = S_3$).

If we do not want to specify the prime p, we may write that an abelian group is *primary* (instead of p-primary).

Theorem 5.13 (Primary Decomposition).

(i) *Every finite abelian group G is a direct sum of its p-primary components:*

$$G = G_{p_1} \oplus \cdots \oplus G_{p_n}.$$

(ii) *Two finite abelian groups G and G' are isomorphic if and only if $G_p \cong G'_p$ for every prime p.*

Proof. (i) Let $x \in G$ be nonzero, and let its order be d. By the fundamental theorem of arithmetic, there are distinct primes $p_1, \ldots, p_n$ and positive exponents $e_1, \ldots, e_n$ with

$$d = p_1^{e_1} \cdots p_n^{e_n}.$$

Define $r_i = d/p_i^{e_i}$, so that $p_i^{e_i} r_i = d$. It follows that $r_i x \in G_{p_i}$ for each i (because $dx = 0$). But the gcd d of $r_1, \ldots, r_n$ is 1 (the only possible prime divisors of d are $p_1, \ldots, p_n$; but no p_i is a common divisor because $p_i \nmid r_i$); hence, there are integers $s_1, \ldots, s_n$ with $1 = \sum_i s_i r_i$. Therefore,

$$x = \sum_i s_i r_i x \in G_{p_1} + \cdots + G_{p_n}.$$

Write $H_i = G_{p_1} + G_{p_2} + \cdots + \widehat{G}_{p_i} + \cdots + G_{p_n}$. By Proposition 5.4, it suffices to prove that if

$$x \in G_{p_i} \cap H_i,$$

then $x = 0$. Since $x \in G_{p_i}$, we have $p_i^\ell x = 0$ for some $\ell \geq 0$; since $x \in H_i$, we have $x = \sum_{j \neq i} y_j$, where $p_j^{g_j} y_j = 0$; hence, $ux = 0$, where $u = \prod_{j \neq i} p_j^{g_j}$. But p_i^ℓ and u are relatively prime, so there exist integers s and t with $1 = sp_i^\ell + tu$. Therefore,

$$x = (sp_i^\ell + tu)x = sp_i^\ell x + tux = 0.$$

(ii) If $f \colon G \to G'$ is a homomorphism, then $f(G_p) \subseteq G'_p$ for every prime p, for if $p^\ell a = 0$, then $0 = f(p^\ell a) = p^\ell f(a)$. If f is an isomorphism, then $f^{-1} \colon G' \to G$ is also an isomorphism [so that $f^{-1}(G'_p) \subseteq G_p$ for all p]. It follows that each restriction $f|G_p \colon G_p \to G'_p$ is an isomorphism, with inverse $f^{-1}|G'_p$.

Conversely, if there are isomorphisms $f_p \colon G_p \to G'_p$ for all p, then there is an isomorphism $\varphi \colon \sum_p G_p \to \sum_p G'_p$ given by $\sum_p a_p \mapsto \sum_p f_p(a_p)$. •

The next type of subgroup will play an important role.

Definition. Let p be a prime and let G be a p-primary abelian group.[1] A subgroup $S \subseteq G$ is a ***pure***[2] ***subgroup*** if, for all $n \geq 0$,

$$S \cap p^n G = p^n S.$$

The inclusion $S \cap p^n G \geq p^n S$ is true for every subgroup $S \subseteq G$, and so it is only the reverse inclusion $S \cap p^n G \subseteq p^n S$ that is significant. It says that if $s \in S$ satisfies an equation $s = p^n a$ for some $a \in G$, then there exists $s' \in S$ with $s = p^n s'$.

Example 5.14.
(i) Every direct summand S of G is a pure subgroup. If $G = S \oplus T$ and $(s, 0) = p^n(u, v)$, where $u \in S$ and $v \in T$, then it is clear that $(s, 0) = p^n(u, 0)$. (The converse: "Every pure subgroup S is a direct summand" is true when S is finite, but it may be false when S is infinite.)

(ii) If $G = \langle a \rangle$ is a cyclic group of order p^2, where p is a prime, then $S = \langle pa \rangle$ is not a pure subgroup of G, for if $s = pa \in S$, then there is no element $s' \in S$ with $s = pa = ps'$. ◀

Lemma 5.15. *If p is a prime and G is a finite p-primary abelian group, then G has a nonzero pure cyclic subgroup.*

Proof. Since G is finite, we may choose an element $y \in G$ of largest order, say, p^ℓ. We claim that $S = \langle y \rangle$ is a pure subgroup of G.

Suppose that $s \in S$, so that $s = mp^t y$, where $t \geq 0$ and $p \nmid m$, and let

$$s = p^n a$$

for some $a \in G$; an element $s' \in S$ must be found with $s = p^n s'$. We may assume that $n < \ell$: otherwise, $s = p^n a = 0$ (for $p^\ell g = 0$ for all $g \in G$ because y has largest order p^ℓ), and we may choose $s' = 0$.

If $t \geq n$, define $s' = mp^{t-n} y \in S$, and note that

$$p^n s' = p^n m p^{t-n} y = mp^t y = s.$$

If $t < n$, then

$$p^\ell a = p^{\ell-n} p^n a = p^{\ell-n} s = p^{\ell-n} m p^t y = m p^{\ell-n+t} y.$$

But $p \nmid m$ and $\ell - n + t < \ell$, because $-n + t < 0$, and so $p^\ell a \neq 0$. This contradicts y having largest order, and so this case cannot occur. •

[1] If G is not a primary group, then a pure subgroup $S \subseteq G$ is defined to be a subgroup that satisfies $S \cap mG = mS$ for all $m \in \mathbb{Z}$ (see Exercises 5.2 and 5.3 on page 267).

[2] Recall that *pure extensions* $k(u)/k$ arose in our discussion of solvability by radicals on page 206; in such an extension, the adjoined element u satisfies the equation $u^n = a$ for some $a \in k$. Pure subgroups are defined in terms of similar equations (written additively), and they are probably so called because of this.

Definition. If p is a prime and G is a finite p-primary abelian group, then

$$d(G) = \dim(G/pG).$$

Observe that d is additive over direct sums,

$$d(G \oplus H) = d(G) + d(H),$$

for Proposition 2.79 gives

$$\begin{aligned}(G \oplus H)/p(G \oplus H) &= (G \oplus H)/(pG \oplus pH)\\ &\cong (G/pG) \oplus (H/pH).\end{aligned}$$

The dimension of the left side is $d(G \oplus H)$ and the dimension of the right-hand side is $d(G) + d(H)$, for the union of a basis of G/pG and a basis of H/pH is a basis of $(G/pG) \oplus (H/pH)$.

The nonzero abelian groups G with $d(G) = 1$ are easily characterized.

Lemma 5.16. *If $G \neq \{0\}$ is p-primary, then $d(G) = 1$ if and only if G is cyclic.*

Proof. If G is cyclic, then so is any quotient of G; in particular, G/pG is cyclic, and so $\dim(G/pG) = 1$.

Conversely, if $G/pG = \langle z + pG \rangle$, then $G/pG \cong \mathbb{I}_p$. Since $\mathbb{I}_p$ is a simple group, the correspondence theorem says that pG is a maximal proper subgroup of G; we claim that pG is the only maximal proper subgroup of G. If $L \subseteq G$ is any maximal proper subgroup, then $G/L \cong \mathbb{I}_p$, for G/L is a simple abelian group of order a power of p, hence has order p (by Proposition 2.107, the abelian simple groups are precisely the cyclic groups of prime order). Thus, if $a \in G$, then $p(a + L) = 0$ in G/L, so that $pa \in L$; hence $pG \subseteq L$. But pG is maximal, and so $pG = L$. It follows that every proper subgroup of G is contained in pG (for every proper subgroup is contained in some maximal proper subgroup). In particular, if $\langle z \rangle$ is a proper subgroup of G, then $\langle z \rangle \subseteq pG$, contradicting $z + pG$ being a generator of G/pG. Therefore, $G = \langle z \rangle$, and so G is cyclic. •

Lemma 5.17. *Let G be a finite p-primary abelian group.*

(i) *If $S \subseteq G$, then $d(G/S) \leq d(G)$.*

(ii) *If S is a pure subgroup of G, then*

$$d(G) = d(S) + d(G/S).$$

Proof. (i) By the correspondence theorem, $p(G/S) = (pG + S)/S$, so that

$$(G/S)/p(G/S) = (G/S)/[(pG + S)/S] \cong G/(pG + S),$$

by the third isomorphism theorem. Since $pG \subseteq pG + S$, there is a surjective homomorphism (of vector spaces over $\mathbb{F}_p$),

$$G/pG \to G/(pG + S),$$

namely, $g + pG \mapsto g + (pG + S)$. Hence, $\dim(G/pG) \geq \dim(G/(pG + S))$; that is, $d(G) \geq d(G/S)$.

(ii) We now analyze $(pG + S)/pG$, the kernel of $G/pG \to G/(pG + S)$. By the second isomorphism theorem,

$$(pG + S)/pG \cong S/(S \cap pG).$$

Since S is a pure subgroup, $S \cap pG = pS$; therefore,

$$(pG + S)/pG \cong S/pS,$$

and so $\dim[(pG + S)/pG] = d(S)$. But if W is a subspace of a finite-dimensional vector space V, then $\dim(V) = \dim(W) + \dim(V/W)$, by Exercise 3.72 on page 170. Hence, if $V = G/pG$ and $W = (pG + S)/pG$, we have

$$d(G) = d(S) + d(G/S). \quad \bullet$$

Theorem 5.18 (Basis Theorem). *Every finite abelian group G is a direct sum of cyclic groups of prime power orders.*

Proof. By the primary decomposition, Theorem 5.13, we may assume that G is p-primary for some prime p. We prove that G is a direct sum of cyclic groups by induction on $d(G) \geq 1$. The base step is easy, for Lemma 5.16 shows that G must be cyclic in this case.

To prove the inductive step, we begin by using Lemma 5.15 to find a nonzero pure cyclic subgroup $S \subseteq G$. By Lemma 5.17, we have

$$d(G/S) = d(G) - d(S) = d(G) - 1 < d(G).$$

By induction, G/S is a direct sum of cyclic groups, say,

$$G/S = \sum_{i=1}^{q} \langle \overline{x}_i \rangle,$$

where $\overline{x}_i = x_i + S$.

Let $x \in G$ and let $\overline{x}$ have order p^ℓ, where $\overline{x} = x + S$. We claim that there is $z \in G$ with $z + S = \overline{x} = x + S$ such that

$$\text{order } z = \text{ order } (\overline{x}).$$

Now x has order p^n, where $n \geq \ell$. But $p^\ell(x + S) = p^\ell \overline{x} = 0$ in G/S, so there is some $s \in S$ with $p^\ell x = s$. By purity, there is $s' \in S$ with $p^\ell x = p^\ell s'$. If we define $z = x - s'$, then $p^\ell z = 0$ and $z + S = x + S = \overline{x}$. If z has order p^m, then $m \geq \ell$ because $z \mapsto \overline{x}$; since $p^\ell z = 0$, the order of z equals p^ℓ.

For each i, choose $z_i \in G$ with $z_i + S = \overline{x}_i = x_i + S$ and with order $z_i =$ order $\overline{x}_i$; define T by

$$T = \langle z_1, \ldots, z_q \rangle.$$

Now $S + T = G$, because G is generated by S and the z_i's. To see that $G = S \oplus T$, it now suffices to prove that $S \cap T = \{0\}$. If $y \in S \cap T$, then $y = \sum_i m_i z_i$, where $m_i \in \mathbb{Z}$. Now $y \in S$, and so $\sum_i m_i \overline{x}_i = 0$ in G/S. Since this is a direct sum, each $m_i \overline{x}_i = 0$; after all, for each i,

$$-m_i \overline{x}_i = \sum_{j \neq i} m_j \overline{x}_j \in \langle \overline{x}_i \rangle \cap \big(\langle \overline{x}_1 \rangle + \cdots + \widehat{\langle \overline{x}_i \rangle} + \cdots + \langle \overline{x}_q \rangle \big) = \{0\}.$$

Therefore, $m_i z_i = 0$ for all i, and hence $y = 0$.

Finally, $G = S \oplus T$ implies $d(G) = d(S) + d(T) = 1 + d(T)$, so that $d(T) < d(G)$. By induction, T is a direct sum of cyclic groups, and this completes the proof. •

The shortest proof of the basis theorem that I know is due to G. Navarro, *American Mathematical Monthly* 110 (2003), pages 153–154.

Lemma 5.19. *A finite p-primary abelian group G is cyclic if and only if it has a unique subgroup of order p.*

Proof. Recall the unnumbered theorem on page 94: If G is an abelian group of order n having at most one cyclic subgroup of order p for every prime divisor p of n, then G is cyclic. The lemma follows at once when n is a power of p. The converse is Lemma 2.85. •

Remark. We cannot remove the hypothesis that G be abelian, for the group $\mathbf{Q}$ of quaternions is a 2-group having a unique subgroup of order 2. However, if G is a (possibly nonabelian) finite p-group having a unique subgroup of order p, then G is either cyclic or generalized quaternion (the latter groups are defined on page 298). A proof of this last result can be found in Rotman, *An Introduction to the Theory of Groups*, pages 121–122.

One cannot remove the finiteness hypothesis, for Proposition 9.25(iii) shows that the infinite p-primary group $\mathbb{Z}(p^\infty)$ has a unique subgroup of order p. ◂

Lemma 5.20. *If G is a finite p-primary abelian group and if a is an element of largest order in G, then $A = \langle a \rangle$ is a direct summand of G.*

Proof. The proof is by induction on $|G| \geq 1$; the base step is trivially true. We may assume that G is not cyclic, for any group is a direct summand of itself (with complementary summand $\{0\}$). Now A has a unique subgroup of order p; call it C. By Lemma 5.19, G contains another subgroup of order p, say C'. Of course, $A \cap C' = \{0\}$. By the second isomorphism theorem, $(A + C')/C' \cong A/(A \cap C') \cong A$ is a cyclic subgroup of G/C'. But no homomorphic image of G can have a cyclic subgroup of order greater than $|A|$ (for no element of an image can have order larger than the order of a). Therefore, $(A + C')/C'$ is a cyclic subgroup of G/C' of largest order and, by the inductive hypothesis, it is a direct summand: There is a subgroup B/C', where $C' \subseteq B \subseteq G$, with

$$G/C' = \big((A + C')/C'\big) \oplus \big(B/C'\big).$$

We claim that $G = A \oplus B$. Clearly, $G = A + C' + B = A + B$ (for $C' \subseteq B$), while $A \cap B \subseteq A \cap \big((A + C') \cap B\big) = A \cap C' = \{0\}$. •

Theorem 5.21 (Basis Theorem). *Every finite abelian group G is a direct sum of cyclic groups.*

Proof. The proof is by induction on $|G| \geq 1$, and the base step is obviously true. To prove the inductive step, let p be a prime divisor of $|G|$. Now $G = G_p \oplus H$, where $p \nmid |H|$ (either we can invoke the primary decomposition or reprove this special case of it). By induction, H is a direct sum of cyclic groups. If G_p is cyclic, we are done. Otherwise, Lemma 5.20 applies to write $G_p = A \oplus B$, where A is cyclic. By the inductive hypothesis, B is a direct sum of cyclic groups, and the theorem is proved. •

Another short proof of the basis theorem is due to R. Rado, *Journal London Mathematical Society* 26 (1951), pages 75–76 and 160. We merely sketch the proof.

Let G be an additive abelian group, and let $x_1, \ldots, x_n$ be elements of G. Form the $1 \times n$ matrix X whose jth entry is x_j. If U is an $n \times n$ matrix with entries in $\mathbb{Z}$, then XU is another $1 \times n$ matrix with entries in G, for its entries are $\mathbb{Z}$-linear combinations of $x_1, \ldots, x_n$. It is easy to check associativity: If U and V are $n \times n$ matrices with entries in $\mathbb{Z}$, then $X(UV) = (XU)V$. Moreover, there is an obvious relation between the subgroups generated by XU and by X; namely, $\langle XU \rangle \subseteq \langle X \rangle$.

Lemma A. *Let G be an additive abelian group, let $x_1, \ldots, x_n$ be elements of G, let X be the $1 \times n$ matrix X whose jth entry is x_j, and let U be an $n \times n$ matrix with entries in $\mathbb{Z}$. If* $\det(U) = 1$, *then $\langle XU \rangle = \langle X \rangle$.*

Definition. An $n \times 1$ matrix $[a_1, \ldots, a_n]$ with entries in a PID R is called a ***unimodular column*** if $\gcd\,(a_1, \ldots, a_n) = 1$.

Lemma B. *If R is a* PID, *then every unimodular column* $[a_1, \ldots, a_n]$ *is the first column of some $n \times n$ matrix U over R with* $\det(U) = 1$.

Sketch of Proof. The proof is by induction on $n \geq 2$. If $n = 2$, then there are elements s and t in R with $ta_1 + sa_2 = 1$, and $U = \begin{bmatrix} a_1 & -s \\ a_2 & t \end{bmatrix}$ is a matrix of determinant 1. The inductive step begins by setting $d = \gcd(a_1, \ldots, a_{n-1})$ and defining $b_i = a_i/d$ for $i \leq n-1$. Since $[b_1, \ldots, b_{n-1}]$ is a unimodular column, the inductive hypothesis says it is the first column of an $(n-1) \times (n-1)$ matrix U' of determinant 1. Now $(a_n, d) = 1$, since $[a_1, \ldots, a_n]$ is a unimodular column, and so there are $s, t \in R$ with $td + sa_n = 1$. These data are used, in a clever way, to modify U' and then augment it to form an $n \times n$ unimodular matrix with first column $[a_1, \ldots, a_n]$. •

Theorem. (i) *If an abelian group $G = \langle x_1, \ldots, x_n \rangle$ and if $[a_1, \ldots, a_n]$ is a unimodular column, then there is a set of n generators of G one of whose elements is $a_1x_1 + \cdots + a_nx_n$.*

(ii)*If $G = \langle x_1, \ldots, x_n \rangle$ is a finite abelian group, then G is a direct sum of cyclic groups.*

Proof. (i) By Lemma B, there is an $n \times n$ matrix U with $\det(U) = 1$ whose first column is $[a_1, \ldots, a_n]$. Since $\det(U) = 1$, Lemma A applies to say that the elements of XU, the first of which is $a_1x_1 + \cdots + a_nx_n$, generate G.

(ii) Let n be the smallest cardinal of any generating set of G, and call such a generating set a *minimal generating set*. The proof is by induction on the number n of elements in a minimal generating set. If $n = 1$, then G is cyclic, and we are done. Of all the elements in minimal generating sets, choose one, say x, having smallest order, say k (so no minimal generating set contains an element of order less than k). Choose a minimal generating set $\{x_1, \ldots, x_{n-1}, x\}$ containing x, and define $x_n = x$. Now $H = \langle x_1, \ldots, x_{n-1}\rangle$ is a proper subgroup of G, by minimality of n, and H is a direct sum of cyclic groups, by the inductive hypothesis. It suffices to prove that $H \cap \langle x_n\rangle = \{0\}$, for then $G = H + \langle x_n\rangle = H \oplus \langle x_n\rangle$, as desired. If, on the contrary, $\langle x_n\rangle \cap H \neq \{0\}$, then there are integers $a_1, \ldots, a_n$ with $a_n x_n \neq 0$ and $a_n x_n = \sum_{i=1}^{n-1} a_i x_i \in H$ (of course, we may assume that $0 < a_n < k$). Let $d = \gcd(a_1, \ldots, a_n)$. Now $[a_1/d, \ldots, a_n/d]$ is a unimodular column, and so the element $g = -(a_n/d)x_n + \sum_{i=1}^{n-1}(a_i/d)x_i$ is part of a minimal generating set of G, by part (i). But $dg = 0$, and so the order of g is a divisor of d; hence, g is an element of a minimal generating set that has order smaller than k, a contradiction. Therefore, $\langle x_n\rangle \cap H = \{0\}$, and so G is a direct sum of cyclic groups. •

Fundamental Theorem

When are two finite abelian groups G and G' isomorphic? By the basis theorem, such groups are direct sums of cyclic groups, and so our first guess is that $G \cong G'$ if they have the same number of cyclic summands of each type. But this hope is dashed by Theorem 2.81, which says that if m and n are relatively prime, then $\mathbb{I}_{mn} \cong \mathbb{I}_m \times \mathbb{I}_n$; for example, $\mathbb{I}_6 \cong \mathbb{I}_2 \times \mathbb{I}_3$. Thus, we retreat and try to count *primary* cyclic summands. But how can we do this? As in the fundamental theorem of arithmetic, we must ask whether there is some kind of unique factorization theorem here.

Before stating the next lemma, recall that we have defined

$$d(G) = \dim(G/pG).$$

In particular, $d(pG) = \dim(pG/p^2G)$ and, more generally,

$$d(p^nG) = \dim(p^nG/p^{n+1}G).$$

Lemma 5.22. *Let G be a finite p-primary abelian group, where p is a prime, and let $G = \sum_j C_j$, where each C_j is cyclic. If $b_n \geq 0$ is the number of summands C_j having order p^n, then there is some $t \geq 1$ with*

$$d(p^nG) = b_{n+1} + b_{n+2} + \cdots + b_t.$$

Proof. Let B_n be the direct sum of all C_j, if any, with order p^n. Thus,

$$G = B_1 \oplus B_2 \oplus \cdots \oplus B_t$$

for some t. Now

$$p^nG = p^nB_{n+1} \oplus \cdots \oplus p^nB_t,$$

because $p^n B_j = \{0\}$ for all $j \le n$. Similarly,

$$p^{n+1}G = p^{n+1}B_{n+2} \oplus \cdots \oplus p^{n+1}B_t.$$

Now Proposition 5.7 shows that $p^n G/p^{n+1}G$ is isomorphic to

$$\big[p^n B_{n+1}/p^{n+1}B_{n+1}\big] \oplus \big[p^n B_{n+2}/p^{n+1}B_{n+2}\big] \oplus \cdots \oplus \big[p^n B_t/p^{n+1}B_t\big].$$

Exercise 5.7 on page 267 gives $d(p^n B_m/p^{n+1}B_m) = \dim(p^n B_m) = b_m$ for all $n < m$; since d is additive over direct sums, we have

$$d(p^n G) = b_{n+1} + b_{n+2} + \cdots + b_t. \quad \bullet$$

The numbers b_n can now be described in terms of G.

Definition. Let G be a finite p-primary abelian group, where p is a prime. For $n \ge 0$, define

$$U_p(n, G) = d(p^n G) - d(p^{n+1}G).$$

Lemma 5.22 shows that

$$d(p^n G) = b_{n+1} + \cdots + b_t$$

and

$$d(p^{n+1}G) = b_{n+2} + \cdots + b_t,$$

so that $U_p(n, G) = b_{n+1}$.

Theorem 5.23. *If p is a prime, then any two decompositions of a finite p-primary abelian group G into direct sums of cyclic groups have the same number of cyclic summands of each type. More precisely, for each $n \ge 0$, the number of cyclic summands having order p^{n+1} is $U_p(n, G)$.*

Proof. By the basis theorem, there exist cyclic subgroups C_i with $G = \sum_i C_i$. The lemma shows, for each $n \ge 0$, that the number of C_i having order p^{n+1} is $U_p(n, G)$, a number that is defined without any mention of the given decomposition of G into a direct sum of cyclics. Thus, if $G = \sum_j D_j$ is another decomposition of G, where each D_j is cyclic, then the number of D_j having order p^{n+1} is also $U_p(n, G)$, as desired. $\bullet$

Corollary 5.24. *If G and G' are finite p-primary abelian groups, then $G \cong G'$ if and only if $U_p(n, G) = U_p(n, G')$ for all $n \ge 0$.*

Proof. If $\varphi : G \to G'$ is an isomorphism, then $\varphi(p^n G) = p^n G'$ for all $n \ge 0$, and so φ induces isomorphisms of the $\mathbb{F}_p$-vector spaces $p^n G/p^{n+1}G \cong p^n G'/p^{n+1}G'$ for all $n \ge 0$ by $p^n g + p^{n+1}G \mapsto p^n \varphi(g) + p^{n+1}G'$. Thus, their dimensions are the same; that is, $U_p(n, G) = U_p(n, G')$.

Conversely, assume that $U_p(n, G) = U_p(n, G')$ for all $n \ge 0$. If $G = \sum_i C_i$ and $G' = \sum_j C'_j$, where the C_i and C'_j are cyclic, then Lemma 5.22 shows that there are the same number of summands of each type, and so it is a simple matter to construct an isomorphism $G \to G'$. $\bullet$

Definition. If G is a p-primary abelian group, then its ***elementary divisors*** are the numbers in the sequence having $U_p(0, G)$ p's, $U_p(1, G)$ p^2's, ..., $U_p(t-1, G)$ p^t's, where p^t is the largest order of a cyclic summand of G.

If G is a finite abelian group, then its ***elementary divisors*** are the elementary divisors of all its primary components.

Theorem 5.25 (Fundamental Theorem of Finite Abelian Groups). *Two finite abelian groups G and G' are isomorphic if and only if they have the same elementary divisors; that is, any two decompositions of G and G' into direct sums of primary cyclic groups have the same number of summands of each order.*

Proof. By the primary decomposition, Theorem 5.13(ii), $G \cong G'$ if and only if, for each prime p, their primary components are isomorphic: $G_p \cong G'_p$. The result now follows from Corollary 5.24. •

Example 5.26.
How many abelian groups are there of order 72? Now $72 = 2^3 3^2$, so that any abelian group of order 72 is the direct sum of groups of order 8 and order 9. There are three groups of order 8, described by the elementary divisors

$$(2, 2, 2), \quad (2, 4), \quad \text{and} \quad (8);$$

there are two groups of order 9, described by the elementary divisors

$$(3, 3) \quad \text{and} \quad (9).$$

Therefore, to isomorphism, there are six abelian groups of order 72. ◂

Here is a second type of decomposition of a finite abelian group into a direct sum of cyclics that does not mention primary groups.

Proposition 5.27. *Every finite abelian group G is a direct sum of cyclic groups*

$$G = S(c_1) \oplus S(c_2) \oplus \cdots \oplus S(c_t),$$

where $t \geq 1$, $S(c_i)$ is a cyclic group of order c_i, and

$$c_1 \mid c_2 \mid \cdots \mid c_t.$$

Proof. Let $p_1, \ldots, p_n$ be the prime divisors of $|G|$. By the basis theorem, we have, for each p_i,

$$G_{p_i} = S(p_i^{e_{i1}}) \oplus S(p_i^{e_{i2}}) \oplus \cdots \oplus S(p_i^{e_{it}}).$$

We may assume that $0 \leq e_{i1} \leq e_{i2} \leq \cdots \leq e_{it}$; moreover, we may allow "dummy" exponents $e_{ij} = 0$ so that the same last index t can be used for all i. Define

$$c_j = p_1^{e_{1j}} p_2^{e_{2j}} \cdots p_n^{e_{nj}}.$$

It is plain that $c_1 \mid c_2 \mid \cdots \mid c_t$. Finally, Theorem 2.81 shows that

$$S(p_1^{e_{1j}}) \oplus S(p_2^{e_{2j}}) \oplus \cdots \oplus S(p_n^{e_{nj}}) \cong S(c_j)$$

for every j. •

Definition. If G is an abelian group, then its ***exponent*** is the smallest positive integer m for which $mG = \{0\}$.

Corollary 5.28. *If G is a finite abelian group and $G = S(c_1) \oplus S(c_2) \oplus \cdots \oplus S(c_t)$, $S(c_i)$ is a cyclic group of order c_i and $c_1 \mid c_2 \mid \cdots \mid c_t$, then c_t is the exponent of G.*

Proof. Since $c_i \mid c_t$ for all i, we have $c_t S(c_i) = 0$ for all i, and so $c_t G = \{0\}$. On the other hand, there is no number e with $1 \leq e < c_t$ with $eS(c_t) = \{0\}$, and so c_t is the smallest positive integer annihilating G. •

Corollary 5.29. *Every noncyclic finite abelian group G has a subgroup isomorphic to $\mathbb{I}_c \oplus \mathbb{I}_c$ for some $c > 1$.*

Proof. By Proposition 5.27, $G = \mathbb{I}_{c_1} \oplus \mathbb{I}_{c_2} \oplus \cdots \oplus \mathbb{I}_{c_t}$, where $t \geq 2$, because G is not cyclic. Since $c_1 \mid c_2$, the cyclic group $\mathbb{I}_{c_2}$ contains a subgroup isomorphic to $\mathbb{I}_{c_1}$, and so G has a subgroup isomorphic to $\mathbb{I}_{c_1} \oplus \mathbb{I}_{c_1}$. •

Let us return to the structure of finite abelian groups.

Definition. If G is a finite abelian group, and if

$$G = S(c_1) \oplus S(c_2) \oplus \cdots \oplus S(c_t),$$

where $t \geq 1$, $S(c_j)$ is a cyclic group of order $c_j > 1$, and $c_1 \mid c_2 \mid \cdots \mid c_t$, then $c_1, c_2, \ldots, c_t$ are called the ***invariant factors*** of G.

Corollary 5.30. *If G is a finite abelian group with invariant factors $c_1, \ldots, c_t$ and elementary divisors $\{p_i^{e_{ij}}\}$, then $|G| = \prod_{j=1}^t c_j = \prod_{ij} p_i^{e_{ij}}$, and its exponent is c_t.*

Proof. We have

$$\begin{aligned} G &\cong \mathbb{Z}/(c_1) \oplus \cdots \oplus \mathbb{Z}/(c_t) \\ &\cong \mathbb{I}_{c_1} \oplus \cdots \oplus \mathbb{I}_{c_t}. \end{aligned}$$

Since the underlying set of a direct sum is the cartesian product, we have $|G| = \prod_{j=1}^t c_j$ and $|G| = \prod_{ij} p_i^{e_{ij}}$. That c_t is the exponent was proved in Corollary 5.28. •

Example 5.31.
In Example 5.26, we displayed the elementary divisors of abelian groups of order 72; here are their invariant factors.

$$\begin{aligned}
\text{elementary divisors} &\leftrightarrow \text{invariant factors}\\
(2,2,2,3,3) = (2,2,2,1,3,3) &\leftrightarrow 2 \mid 6 \mid 6\\
(2,4,3,3) &\leftrightarrow 6 \mid 12\\
(8,3,3) = (1,8,3,3) &\leftrightarrow 3 \mid 24\\
(2,2,2,9) = (2,2,2,1,1,9) &\leftrightarrow 2 \mid 2 \mid 18\\
(2,4,9) = (2,4,1,9) &\leftrightarrow 2 \mid 36\\
(8,9) &\leftrightarrow 72 \quad \blacktriangleleft
\end{aligned}$$

Theorem 5.32 (Invariant Factors). *Two finite abelian groups are isomorphic if and only they have the same invariant factors.*

Proof. Given the elementary divisors of G, we can construct invariant factors, as in the proof of Proposition 5.27:

$$c_j = p_1^{e_{1j}} p_2^{e_{2j}} \cdots p_n^{e_{nj}},$$

where those factors $p_i^{e_{i1}}, p_i^{e_{i2}}, \cdots$ not equal to $p_i^0 = 1$ are the elementary divisors of the p_i-primary component of G. Thus, the invariant factors depend only on G because they are defined in terms of the elementary divisors.

To prove isomorphism, it suffices, by the fundamental theorem, to prove that the elementary divisors can be computed from the invariant factors. Since $c_j = p_1^{e_{1j}} p_2^{e_{2j}} \cdots p_n^{e_{nj}}$, the fundamental theorem of arithmetic shows that c_j determines all those prime powers $p_i^{e_{ij}}$ which are distinct from 1; that is, the invariant factors c_j determine the elementary divisors. •

In Example 5.31, we started with elementary divisors and computed invariant factors. Let us now start with invariant factors and compute elementary divisors.

$$\begin{aligned}
\text{invariant factors} &\leftrightarrow \text{elementary divisors}\\
2 \mid 6 \mid 6 = 2 \mid 2\cdot 3 \mid 2 \cdot 3 &\leftrightarrow (2,2,2,3,3)\\
6 \mid 12 = 2 \cdot 3 \mid 2^2 \cdot 3 &\leftrightarrow (2,4,3,3)\\
3 \mid 24 = 3 \mid 2^3 \cdot 3 &\leftrightarrow (8,3,3)\\
2 \mid 2 \mid 18 = 2 \mid 2 \mid 2 \cdot 3^2 &\leftrightarrow (2,2,2,9)\\
2 \mid 36 = 2 \mid 2^2 \cdot 3^2 &\leftrightarrow (2,4,9)\\
72 = 2^3 \cdot 3^2 &\leftrightarrow (8,9).
\end{aligned}$$

The results of this section will be generalized, in Chapter 9, from finite abelian groups to finitely generated abelian groups, where an abelian group G is *finitely generated* if there are finitely many elements $a_1, \ldots, a_n \in G$ so that every $x \in G$ is a linear combination

of them: $x = \sum_i m_i a_i$, where $m_i \in \mathbb{Z}$ for all i. The basis theorem generalizes: Every finitely generated abelian group G is a direct sum of cyclic groups, each of which is a finite primary group or an infinite cyclic group; the fundamental theorem also generalizes: Given two decompositions of G into a direct sum of cyclic groups (as in the basis theorem), the number of cyclic summands of each type is the same in both decompositions. The basis theorem is no longer true for abelian groups that are not finitely generated; for example, the additive group $\mathbb{Q}$ of rational numbers is not a direct sum of cyclic groups.

EXERCISES

5.1 (i) Let G be an arbitrary, possibly nonabelian, group, and let S and T be normal subgroups of G. Prove that if $S \cap T = \{1\}$, then $st = ts$ for all $s \in S$ and $t \in T$.
Hint. Show that $sts^{-1}t^{-1} \in S \cap T$.
(ii) Prove that Proposition 5.4 holds for nonabelian groups G if we assume that all the subgroups S_i are normal subgroups.

5.2 Let G be an abelian group, not necessarily primary. Define a subgroup $S \subseteq G$ to be a ***pure subgroup*** if, for all $m \in \mathbb{Z}$,

$$S \cap mG = mS.$$

Prove that if G is a p-primary abelian group, then a subgroup $S \subseteq G$ is pure as just defined if and only if $S \cap p^n G = p^n S$ for all $n \geq 0$ (the definition in the text).

5.3 Let G be a possibly infinite abelian group.
(i) Prove that every direct summand S of G is a pure subgroup.
Define the ***torsion***[3] ***subgroup*** tG of G as

$$tG = \{a \in G : a \text{ has finite order}\}.$$

(ii) Prove that tG is a pure subgroup of G. [There exist abelian groups G whose torsion subgroup tG is not a direct summand (see Exercise 9.1(iii) on page 663); hence, a pure subgroup need not be a direct summand.]
(iii) Prove that G/tG is an abelian group in which every nonzero element has infinite order.

5.4 Let p be a prime and let q be relatively prime to p. Prove that if G is a p-group and $g \in G$, then there exists $x \in G$ with $qx = g$.

5.5 Let $G = \langle a \rangle$ be a cyclic group of finite order m. Prove that G/nG is a cyclic group of order d, where $d = (m, n)$.

5.6 For a group G and a positive integer n, define

$$G[n] = \{g \in G : g^n = 1\}.$$

Prove that $G[n] = \langle a^{m/d} \rangle$, where $d = (m, n)$, and conclude that $G[n] \cong \mathbb{I}_d$.

5.7 Prove that if $B = B_m = \langle x_1 \rangle \oplus \cdots \oplus \langle x_{b_m} \rangle$ is a direct sum of b_m cyclic groups of order p^m, and if $n < m$, then the cosets $p^n x_i + p^{n+1}B$, for $1 \leq i \leq b_m$ are a basis for $p^n B / p^{n+1} B$. Conclude that $d(p^n B_m) = b_m$ when $n < m$. [Recall that if G is a finite abelian group, then G/pG is a vector space over $\mathbb{F}_p$ and $d(G) = \dim(G/pG)$.]

[3] This terminology comes from algebraic topology. To each space X, a sequence of abelian groups is assigned, called *homology groups*, and if X is "twisted," then there are elements of finite order in some of these groups.

5.8 (i) If G is a finite p-primary abelian group, where p is a prime, and if $x \in G$ has largest order, prove that $\langle x \rangle$ is a direct summand of G.

(ii) Prove that if G is a finite abelian group and $x \in G$ has maximal order (that is, there is no element in G having larger order), then $\langle x \rangle$ is a direct summand of G.

5.9 Prove that a subgroup of a finite abelian group is a direct summand if and only if it is a pure subgroup.

Hint. Modify the proof of the basis theorem, Theorem 5.18.

5.10 (i) If G and H are finite abelian groups, prove, for all primes p and all $n \geq 0$, that

$$U_p(n, G \oplus H) = U_p(n, G) + U_p(n, H).$$

(ii) If A, B, and C are finite abelian groups, prove that $A \oplus B \cong A \oplus C$ implies $B \cong C$.

(iii) If A and B are finite abelian groups, prove that $A \oplus A \cong B \oplus B$ implies $A \cong B$.

5.11 If n is a positive integer, then a ***partition of*** n is a sequence of positive integers $i_1 \leq i_2 \leq \cdots \leq i_r$ with $i_1 + i_2 + \cdots + i_r = n$. If p is a prime, prove that the number of nonisomorphic abelian groups of order p^n is equal to the number of partitions of n.

5.12 Prove that there are, to isomorphism, exactly 14 abelian groups of order 288.

5.13 Prove the uniqueness assertion in the fundamental theorem of arithmetic by applying the fundamental theorem of finite abelian groups to $G = \mathbb{I}_n$.

5.14 (i) If G is a finite abelian group, define

$$\nu_k(G) = \text{ the number of elements in } G \text{ of order } k.$$

Prove that two finite abelian groups G and G' are isomorphic if and only if $\nu_k(G) = \nu_k(G')$ for all integers k.

Hint. If B is a direct sum of k copies of a cyclic group of order p^n, then how many elements of order p^n are in B?

(ii) Give an example of two nonisomorphic not necessarily abelian finite groups G and G' for which $\nu_k(G) = \nu_k(G')$ for all integers k.

Hint. Take G of order p^3.

5.15 Prove that the additive group $\mathbb{Q}$ is not a direct sum: $\mathbb{Q} \ncong A \oplus B$, where A and B are nonzero subgroups.

Hint. If $a, b \in \mathbb{Q}$ are not zero, then there is $c \in \mathbb{Q}$ with $a, b \in \langle c \rangle$.

5.16 Let $G = B_1 \oplus B_2 \oplus \cdots \oplus B_t$, where the B_i are subgroups.

(i) Prove that $G[p] = B_1[p] \oplus B_2[p] \oplus \cdots \oplus B_t[p]$.

(ii) Prove, for all $n \geq 0$ that

$$\begin{aligned} p^n G \cap G[p] &= \big(p^n G \cap B_1[p]\big) \oplus \big(p^n G \cap B_2[p]\big) \oplus \cdots \oplus \big(p^n G \cap B_t[p]\big) \\ &= \big(p^n B_1 \cap B_1[p]\big) \oplus \big(p^n B_2 \cap B_2[p]\big) \oplus \cdots \oplus \big(p^n B_t \cap B_t[p]\big). \end{aligned}$$

(iii) If G is a p-primary abelian group, prove, for all $n \geq 0$, that

$$U_p(n, G) = \dim \left(\frac{p^n G \cap G[p]}{p^{n+1} G \cap G[p]} \right).$$

5.2 THE SYLOW THEOREMS

We return to nonabelian groups, and so we revert to the multiplicative notation. The Sylow theorems are analogs, for finite nonabelian groups, of the primary components of finite abelian groups.

Recall that a group $G \neq \{1\}$ is called *simple* if it has no normal subgroups other than $\{1\}$ and G itself. We saw, in Proposition 2.107, that the abelian simple groups are precisely the cyclic groups $\mathbb{I}_p$ of prime order p, and we saw, in Theorem 2.112, that A_n is a nonabelian simple group for all $n \geq 5$. In fact, A_5 is the nonabelian simple group of smallest order. How can we prove that a nonabelian group G of order less than $60 = |A_5|$ is not simple? Exercise 2.98 on page 114 states that if G is a group of order $|G| = mp$, where p is prime and $1 < m < p$, then G is not simple. This exercise shows that many of the numbers less than 60 are not orders of simple groups. After throwing out all prime powers (p-groups are never nonabelian simple), the only remaining possibilities are

$$12, 18, 24, 30, 36, 40, 45, 48, 50, 54, 56.$$

The solution to the exercise uses Cauchy's theorem, which says that G has a subgroup of order p. We shall see that if G has a subgroup of order p^e instead of p, then Exercise 2.98 can be generalized, and the list of candidates can be shortened. What proper subgroups of G do we know other than cyclic subgroups? The center $Z(G)$ of a group G is a possible candidate, but this subgroup might not be proper or it might be trivial: if G is abelian, then $Z(G) = G$; if $G = S_3$, then $Z(G) = \{1\}$. Hence, $Z(G)$ cannot be used to generalize the exercise.

Traité des Substitutions et des Équations Algébriques, by C. Jordan, published in 1870, was the first book on group theory (more than half of it is devoted to Galois theory, then called the theory of equations). At about the same time, but too late for publication in Jordan's book, three fundamental theorems were discovered. In 1868, E. Schering proved the basis theorem: Every finite abelian group is a direct product of cyclic groups, each of prime power order; in 1870, L. Kronecker, unaware of Schering's proof, also proved this result. In 1878, F. G. Frobenius and L. Stickelberger proved the fundamental theorem of finite abelian groups. In 1872, L. Sylow showed, for every finite group G and every prime p, that if p^e is the largest power of p dividing $|G|$, then G has a subgroup of order p^e, nowadays called a *Sylow subgroup*; we will use such subgroups to generalize Exercise 2.98. Our strategy for proving the Sylow theorems works best if we adopt the following definition.

Definition. Let p be a prime. A ***Sylow p-subgroup*** of a finite group G is a maximal p-subgroup P.

Maximality means that if Q is a p-subgroup of G and $P \leq Q$, then $P = Q$.

It follows from Lagrange's theorem that if p^e is the largest power of p dividing $|G|$, then a subgroup of order p^e, should it exist, is a maximal p-subgroup of G. One virtue of the present definition is that maximal p-subgroups always exist: indeed, we now show

that if S is any p-subgroup of G (perhaps $S = \{1\}$), then there exists a Sylow p-subgroup P containing S. If there is no p-subgroup strictly containing S, then S itself is a Sylow p-subgroup. Otherwise, there is a p-subgroup P_1 with $S < P_1$. If P_1 is maximal, it is Sylow, and we are done. Otherwise, there is some p-subgroup P_2 with $P_1 < P_2$. This procedure of producing larger and larger p-subgroups P_i must end after a finite number of steps, for $|P_i| \leq |G|$ for all i; the largest P_i must, therefore, be a Sylow p-subgroup.

Recall that a *conjugate* of a subgroup $H \leq G$ is a subgroup of G of the form

$$aHa^{-1} = \{aha^{-1} : h \in H\},$$

where $a \in G$. The *normalizer* of H in G is the subgroup

$$N_G(H) = \{a \in G : aHa^{-1} = H\},$$

and Proposition 2.101 states that if H is a subgroup of a finite group G, then the number of conjugates of H in G is $[G : N_G(H)]$.

It is obvious that $H \lhd N_G(H)$, and so the quotient group $N_G(H)/H$ is defined.

Lemma 5.33. *Let P be a Sylow p-subgroup of a finite group G.*

(i) *Every conjugate of P is also a Sylow p-subgroup of G.*

(ii) *$|N_G(P)/P|$ is prime to p.*

(iii) *If $a \in G$ has order some power of p and if $aPa^{-1} = P$, then $a \in P$.*

Proof. (i) If $a \in G$, then aPa^{-1} is a p-subgroup of G; if it is not a maximal p-subgroup, then there is a p-subgroup Q with $aPa^{-1} < Q$. Hence, $P < a^{-1}Qa$, contradicting the maximality of P.

(ii) If p divides $|N_G(P)/P|$, then Cauchy's theorem shows that $N_G(P)/P$ contains an element aP of order p, and hence $N_G(P)/P$ contains a subgroup $S^* = \langle aP \rangle$ of order p. By the correspondence theorem (Theorem 2.76), there is a subgroup S with $P \leq S \leq N_G(P)$ such that $S/P \cong S^*$. But S is a p-subgroup of $N_G(P) \leq G$ (by Exercise 2.75 on page 95) strictly larger than P, and this contradicts the maximality of P. We conclude that p does not divide $|N_G(P)/P|$.

(iii) By the definition of normalizer, the element a lies in $N_G(P)$. If $a \notin P$, then the coset aP is a nontrivial element of $N_G(P)/P$ having order some power of p; in light of part (ii), this contradicts Lagrange's theorem. •

Since every conjugate of a Sylow p-subgroup is a Sylow p-subgroup, it is reasonable to let G act by conjugation on the Sylow p-subgroups.

Theorem 5.34 (*Sylow*). *Let G be a finite group of order $p_1^{e_1} \cdots p_t^{e_t}$, and let P be a Sylow p-subgroup of G for some prime $p = p_j$.*

(i) *Every Sylow p-subgroup is conjugate to P.*

(ii) *If there are r_j Sylow p_j-subgroups, then r_j is a divisor of $|G|/p_j^{e_j}$ and*

$$r_j \equiv 1 \bmod p_j.$$

Proof. Let $X = \{P_1, \ldots, P_{r_j}\}$ be the set of all the conjugates of P, where we have denoted P by P_1. If Q is any Sylow p-subgroup of G, then Q acts on X by conjugation: If $a \in Q$, then it sends

$$P_i = g_i P g_i^{-1} \mapsto a\big(g_i P g_i^{-1}\big)a^{-1} = (ag_i)P(ag_i)^{-1} \in X.$$

By Corollary 2.99, the number of elements in any orbit is a divisor of $|Q|$; that is, every orbit has size some power of p (because Q is a p-group). If there is an orbit of size 1, then there is some P_i with $aP_ia^{-1} = P_i$ for all $a \in Q$. By Lemma 5.33, we have $a \in P_i$ for all $a \in Q$; that is, $Q \leq P_i$. But Q, being a Sylow p-subgroup, is a maximal p-subgroup of G, and so $Q = P_i$. In particular, if $Q = P_1$, then there is only one orbit of size 1, namely, $\{P_1\}$, and all the other orbits have sizes that are honest powers of p. We conclude that $|X| = r_j \equiv 1 \bmod p$.

Suppose now that there is some Sylow p-subgroup Q that is not a conjugate of P; thus, $Q \neq P_i$ for any i. Again, we let Q act on X, and again, we ask if there is an orbit of size 1, say, $\{P_k\}$. As in the previous paragraph, this implies $Q = P_k$, contrary to our present assumption that $Q \notin X$. Hence, there are no orbits of size 1, which says that each orbit has size an honest power of p. It follows that $|X| = r_j$ is a multiple of p; that is, $r_j \equiv 0 \bmod p$, which contradicts the congruence $r_j \equiv 1 \bmod p$. Therefore, no such Q can exist, and so all Sylow p-subgroups are conjugate to P.

Finally, since all Sylow p-subgroups are conjugate, we have $r_j = [G : N_G(P)]$, and so r_j is a divisor of $|G|$. But $r_j \equiv 1 \bmod p_j$ implies $(r_j, p_j^{e_j}) = 1$, so that Euclid's lemma gives $r_j \mid |G|/p_j^{e_j}$. •

Corollary 5.35. *A finite group G has a unique Sylow p-subgroup P, for some prime p, if and only if $P \lhd G$.*

Proof. Assume that P, a Sylow p-subgroup of G, is unique. For each $a \in G$, the conjugate aPa^{-1} is also a Sylow p-subgroup; by uniqueness, $aPa^{-1} = P$ for all $a \in G$, and so $P \lhd G$.

Conversely, assume that $P \lhd G$. If Q is any Sylow p-subgroup, then $Q = aPa^{-1}$ for some $a \in G$; but $aPa^{-1} = P$, by normality, and so $Q = P$. •

The following result gives the order of a Sylow subgroup.

Theorem 5.36 (*Sylow*). *If G is a finite group of order $p^e m$, where p is a prime and $p \nmid m$, then every Sylow p-subgroup P of G has order p^e.*

Proof. We first show that $p \nmid [G : P]$. Now

$$[G : P] = [G : N_G(P)][N_G(P) : P].$$

The first factor, $[G : N_G(P)] = r$, is the number of conjugates of P in G, and so p does not divide $[G : N_G(P)]$ because $r \equiv 1 \bmod p$. The second factor, $[N_G(P) : P] = |N_G(P)/P|$, is also not divisible by p, by Lemma 5.33. Therefore, p does not divide $[G : P]$, by Euclid's lemma.

Now $|P| = p^k$ for some $k \leq e$, and so

$$[G : P] = |G|/|P| = p^e m/p^k = p^{e-k}m.$$

Since p does not divide $[G : P]$, we must have $k = e$; that is, $|P| = p^e$. •

Example 5.37.
(i) Let $G = S_4$. Now $|S_4| = 24 = 2^3 3$. Thus, a Sylow 2-subgroup of S_4 has order 8. We have seen, in Exercise 2.83 on page 113, that S_4 contains a copy of the dihedral group D_8 consisting of the symmetries of a square. The Sylow theorem says that all subgroups of order 8 are conjugate, hence isomorphic, to D_8. Moreover, the number r of Sylow 2-subgroups is a divisor of 24 congruent to 1 mod 2; that is, r is an odd divisor of 24. Since $r \neq 1$ (see Exercise 5.17 on page 277), there are exactly three Sylow 2-subgroups.

(ii) If G is a finite abelian group, then a Sylow p-subgroup is just its p-primary component (since G is abelian, every subgroup is normal, and so there is a unique Sylow p-subgroup for every prime p). ◂

Here is a second proof of the last Sylow theorem, due to H. Wielandt.

Theorem 5.38. *If G is a finite group of order $p^e m$, where p is a prime and $p \nmid m$, then G has a subgroup of order p^e.*

Proof. If X is the family of all those subsets of G having exactly p^e elements, then $|X| = \binom{p^e m}{p^e}$; by Exercise 1.29 on page 14, $p \nmid |X|$. Now G acts on X: define gB, for $g \in G$ and $B \in X$, by

$$gB = \{gb : b \in B\}.$$

If p divides $|\mathcal{O}(B)|$ for every $B \in X$, where $\mathcal{O}(B)$ is the orbit of B, then p is a divisor of $|X|$, for X is the disjoint union of orbits, by Proposition 2.97. As $p \nmid |X|$, there exists a subset B with $|B| = p^e$ and with $|\mathcal{O}(B)|$ not divisible by p. If G_B is the stabilizer of this subset B, then Theorem 2.98 gives $[G : G_B] = |\mathcal{O}(B)|$, and so $|G| = |G_B| \cdot |\mathcal{O}(B)|$. Since $p^e \mid |G|$ and $p \nmid \mathcal{O}(B)|$, repeated application of Euclid's lemma gives $p^e \mid |G_B|$. Therefore, $p^e \leq |G_B|$.

For the reverse inequality, choose an element $b \in B$ and define a function $\tau : G_B \to B$ by $g \mapsto gb$. Note that $\tau(g) = gb \in gB = B$, for $g \in G_B$, the stabilizer of B. If g, $h \in G_B$ and $h \neq g$, then $\tau(h) = hb \neq gb = \tau(g)$; that is, τ is an injection. We conclude that $|G_B| \leq |B| = p^e$, and so G_B is a subgroup of G of order p^e. •

Proposition 5.39. *A finite group G all of whose Sylow subgroups are normal is the direct product of its Sylow subgroups.*

Proof. Let $|G| = p_1^{e_1} \cdots p_t^{e_t}$ and let G_{p_i} be the Sylow p_i-subgroup of G. We use Exercise 5.1 on page 267, the generalization of Proposition 5.4 to nonabelian groups. The subgroup S generated by all the Sylow subgroups is G, for $p_i^{e_i} \mid |S|$ for all i. Finally, if $x \in G_{p_i} \bigcap \langle \bigcup_{j \neq i} G_{p_j} \rangle$, then $x = s_i \in G_{p_i}$ and $x = \prod_{j \neq i} s_j$, where $s_j \in G_{p_j}$. Now $x^{p_i^n} = 1$ for some $n \geq 1$. On the other hand, there is some power of p_j, say q_j, with $s_j^{q_j} = 1$ for all j. Since the s_j commute with each other, by Exercise 5.1 on page 267, we have $1 = x^q = (\prod_{j \neq i} s_j)^q$, where $q = \prod_{j \neq i} q_j$. Since $(p_i^n, q) = 1$, there are integers u and v with $1 = up_i^n + vq$, and so $x = x^1 = x^{up_i^n} x^{vq} = 1$. Therefore, G is the direct product of its Sylow subgroups. •

We can now generalize Exercise 2.98 on page 114 and its solution.

Lemma 5.40. *There is no nonabelian simple group G of order $|G| = p^e m$, where p is prime, $p \nmid m$, and $p^e \nmid (m-1)!$.*

Proof. We claim that if p is a prime, then every p-group G with $|G| > p$ is not simple. Theorem 2.75 says that the center, $Z(G)$, is nontrivial. But $Z(G) \lhd G$, so that if $Z(G)$ is a proper subgroup, then G is not simple. If $Z(G) = G$, then G is abelian, and Proposition 2.78 says that G is not simple unless $|G| = p$.

Suppose that such a simple group G exists. By Sylow's theorem, G contains a subgroup P of order p^e, hence of index m. We may assume that $m > 1$, for nonabelian p-groups are never simple. By Theorem 2.88, there exists a homomorphism $\varphi \colon G \to S_m$ with $\ker \varphi \leq P$. Since G is simple, however, it has no proper normal subgroups; hence $\ker \varphi = \{1\}$ and φ is an injection; that is, $G \cong \varphi(G) \leq S_m$. By Lagrange's theorem, $p^e m \mid m!$, and so $p^e \mid (m-1)!$, contrary to the hypothesis. •

Proposition 5.41. *There are no nonabelian simple groups of order less than* 60.

Proof. The reader may now check that the only integers n between 2 and 59, neither a prime power nor having a factorization of the form $n = p^e m$ as in the statement of the lemma, are $n = 30, 40$, and 56. By the lemma, these three numbers are the only candidates for orders of nonabelian simple groups of order < 60.

Assume there is a simple group G of order 30. Let P be a Sylow 5-subgroup of G, so that $|P| = 5$. The number r_5 of conjugates of P is a divisor of 30 and $r_5 \equiv 1 \bmod 5$. Now $r_5 \neq 1$ lest $P \lhd G$, so that $r_5 = 6$. By Lagrange's theorem, the intersection of any two of these is trivial (intersections of Sylow subgroups can be more complicated; see Exercise 5.18 on page 277). There are four nonidentity elements in each of these subgroups, and so there are $6 \times 4 = 24$ nonidentity elements in their union. Similarly, the number r_3 of Sylow 3-subgroups of G is 10 (for $r_3 \neq 1$, r_3 is a divisor of 30, and $r_3 \equiv 1 \bmod 3$). There are two nonidentity elements in each of these subgroups, and so the union of these subgroups has 20 nonidentity elements. We have exceeded the number of elements in G, and so G cannot be simple.

Let G be a group of order 40, and let P be a Sylow 5-subgroup of G. If r is the number of conjugates of P, then $r \mid 40$ and $r \equiv 1 \bmod 5$. These conditions force $r = 1$, so that $P \lhd G$; therefore, no simple group of order 40 can exist.

Finally, assume there is a simple group G of order 56. If P is a Sylow 7-subgroup of G, then P must have $r_7 = 8$ conjugates (for $r_7 \mid 56$ and $r_7 \equiv 1 \bmod 7$). Since these groups are cyclic of prime order, the intersection of any pair of them is $\{1\}$, and so there are 48 nonidentity elements in their union. Thus, adding the identity, we have accounted for 49 elements of G. Now a Sylow 2-subgroup Q has order 8, and so it contributes seven more nonidentity elements, giving 56 elements. But there is a second Sylow 2-subgroup, lest $Q \lhd G$, and we have exceeded our quota. Therefore, there is no simple group of order 56. •

The "converse" of Lagrange's theorem is false: If G is a finite group of order n, and if $d \mid n$, then G may not have a subgroup of order d. For example, we proved, in Proposition 2.64, that the alternating group A_4 is a group of order 12 having no subgroup of order 6.

Proposition 5.42. *Let G be a finite group. If p is a prime and if p^k divides $|G|$, then G has a subgroup of order p^k.*

Proof. If $|G| = p^e m$, where $p \nmid m$, then a Sylow p-subgroup P of G has order p^e. Hence, if p^k divides $|G|$, then p^k divides $|P|$. By Proposition 2.106, P has a subgroup of order p^k; a fortiori, G has a subgroup of order p^k. •

What examples of p-groups have we seen? Of course, cyclic groups of order p^n are p-groups, as is any direct product of copies of these. By the fundamental theorem, this describes all (finite) abelian p-groups. The only nonabelian examples we have seen so far are the dihedral groups D_{2n} (which are 2-groups when n is a power of 2) and the quaternions $\mathbf{Q}$ of order 8 (of course, for every 2-group A, the direct products $D_8 \times A$ and $\mathbf{Q} \times A$ are also nonabelian 2-groups). Here are some new examples.

Definition. A ***unitriangular*** matrix over a field k is an upper triangular matrix each of whose diagonal terms is 1. Define $\mathrm{UT}(n, k)$ to be the set of all $n \times n$ unitriangular matrices over k.

Remark. We can generalize this definition by allowing k to be any commutative ring. For example, the group $\mathrm{UT}(n, \mathbb{Z})$ is an interesting group. ◂

Proposition 5.43. *If k is a field, then $\mathrm{UT}(n, k)$ is a subgroup of $\mathrm{GL}(n, k)$.*

Proof. Of course, the identity I is unitriangular, so that $I \in \mathrm{UT}(n, k)$. If $A \in \mathrm{UT}(n, k)$, then $A = I + N$, where N is *strictly* upper triangular; that is, N is an upper triangular matrix having only 0's on its diagonal. Note that the sum and product of strictly upper triangular matrices is again strictly upper triangular.

Let $e_1, \ldots, e_n$ be the standard basis of k^n. If N is strictly upper triangular, define $T : k^n \to k^n$ by $T(e_i) = Ne_i$, where e_i is regarded as a column matrix. Now T satisfies the equations, for all i,

$$T(e_1) = 0 \quad \text{and} \quad T(e_{i+1}) \in \langle e_1, \ldots, e_i \rangle.$$

It is easy to see, by induction on i, that

$$T^i(e_j) = 0 \text{ for all } j \leq i.$$

It follows that $T^n = 0$ and, hence, that $N^n = 0$. Thus, if $A \in \mathrm{UT}(n, k)$, then $A = I + N$, where $N^n = 0$.

To see that $\mathrm{UT}(n, k)$ is a subgroup of $\mathrm{GL}(n, k)$, first note that $(I + N)(I + M) = I + (N + M + NM)$ is unitriangular. Second, we show that if A is unitriangular, then it is nonsingular and that its inverse is also unitriangular. In analogy to the power series expansion $1/(1 + x) = 1 - x + x^2 - x^3 + \cdots$, we define the inverse of $A = I + N$ to be $B = I - N + N^2 - N^3 + \cdots$ (note that this series stops after $n - 1$ terms because $N^n = 0$), The reader may now check that $BA = I = AB$, so that $B = A^{-1}$. Moreover, N strictly upper triangular implies that $-N + N^2 - N^3 + \cdots \pm N^{n-1}$ is also strictly upper triangular, and so A^{-1} is unitriangular. (Alternatively, for readers familiar with linear algebra, we know that A is nonsingular, because its determinant is 1, and the formula for A^{-1} in terms of its adjoint [the matrix of cofactors] shows that A^{-1} is unitriangular.) Hence, $\mathrm{UT}(n, k)$ is a subgroup of $\mathrm{GL}(n, k)$. •

Proposition 5.44. *Let $q = p^e$, where p is a prime. For each $n \geq 2$, $\mathrm{UT}(n, \mathbb{F}_q)$ is a p-group of order $q^{\binom{n}{2}} = q^{n(n-1)/2}$.*

Proof. The number of entries in an $n \times n$ unitriangular matrix lying strictly above the diagonal is $\binom{n}{2} = \frac{1}{2}n(n-1)$ (throw away n diagonal entries from the total of n^2 entries; half of the remaining $n^2 - n$ entries are above the diagonal). Since each of these entries can be any element of $\mathbb{F}_q$, there are exactly $q^{\binom{n}{2}}$ $n \times n$ unitriangular matrices over $\mathbb{F}_q$, and so this is the order of $\mathrm{UT}(n, \mathbb{F}_q)$. •

Recall Exercise 2.26 on page 62: If G is a group and $x^2 = 1$ for all $x \in G$, then G is abelian. We now ask whether a group G satisfying $x^p = 1$ for all $x \in G$, where p is an odd prime, must also be abelian.

Proposition 5.45. *If p is an odd prime, then there exists a nonabelian group G of order p^3 with $x^p = 1$ for all $x \in G$.*

Proof. If $G = \mathrm{UT}(3, \mathbb{F}_p)$, then $|G| = p^3$. Now G is not abelian; for example, the matrices

$$\begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix}$$

do not commute. If $A \in G$, then $A = I + N$; since p is an odd prime, $p \geq 3$, and so $N^p = 0$. The set of all matrices of the form $a_0 I + a_1 N + \cdots + a_m N^m$, where $a_i \in \mathbb{F}_p$, is easily seen to be a commutative ring in which $pM = 0$ for all M. But Proposition 3.2(vi) says that the binomial theorem holds in every commutative ring; since $p \mid \binom{p}{i}$ when $1 < i < p$, by Proposition 1.12, we have

$$A^p = (I + N)^p = I^p + N^p = I. \quad \bullet$$

Theorem 5.46. *Let $\mathbb{F}_q$ denote the finite field with q elements. Then*

$$|\mathrm{GL}(n, \mathbb{F}_q)| = (q^n - 1)(q^n - q)(q^n - q^2) \cdots (q^n - q^{n-1}).$$

Proof. Let V be an n-dimensional vector space over $\mathbb{F}_q$. We show first that there is a bijection $\Phi\colon \mathrm{GL}(n, \mathbb{F}_q) \to \mathcal{B}$, where $\mathcal{B}$ is the set of all bases of V. Choose, once for all, a basis $e_1, \ldots, e_n$ of V. If $T \in \mathrm{GL}(n, \mathbb{F}_q)$, define

$$\Phi(T) = Te_1, \ldots, Te_n.$$

By Lemma 3.103, $\Phi(T) \in \mathcal{B}$ because T, being nonsingular, carries a basis into a basis. But Φ is a bijection, for given a basis $v_1, \ldots, v_n$, there is a unique linear transformation S, necessarily nonsingular (by Lemma 3.103), with $Se_i = v_i$ for all i (by Theorem 3.92).

Our problem now is to count the number of bases $v_1, \ldots, v_n$ of V. There are q^n vectors in V, and so there are $q^n - 1$ candidates for v_1 (the zero vector is not a candidate). Having chosen v_1, we see that the candidates for v_2 are those vectors not in $\langle v_1 \rangle$, the subspace spanned by v_1; there are thus $q^n - q$ candidates for v_2. More generally, having chosen a linearly independent list $v_1, \ldots, v_i$, then v_{i+1} can be any vector not in $\langle v_1, \ldots, v_i \rangle$. Thus, there are $q^n - q^i$ candidates for v_{i+1}. The result follows by induction on n. $\bullet$

Theorem 5.47. *If p is a prime and $q = p^m$, then the unitriangular group $\mathrm{UT}(n, \mathbb{F}_q)$ is a Sylow p-subgroup of $\mathrm{GL}(n, \mathbb{F}_q)$.*

Proof. Since $q^n - q^i = q^i(q^{n-i} - 1)$, the highest power of p dividing $|\,\mathrm{GL}(n, \mathbb{F}_q)|$ is

$$qq^2q^3 \cdots q^{n-1} = q^{\binom{n}{2}}.$$

But $|\mathrm{UT}(n, \mathbb{F}_q)| = q^{\binom{n}{2}}$, and so it must be a Sylow p-subgroup. $\bullet$

Corollary 5.48. *If p is a prime and G is a finite p-group, then G is isomorphic to a subgroup of the unitriangular group $\mathrm{UT}(|G|, \mathbb{F}_p)$.*

Proof. We show first, for every $m \geq 1$, that the symmetric group S_m can be imbedded in $\mathrm{GL}(m, k)$, where k is a field. Let V be an m-dimensional vector space over k, and let $v_1, \ldots, v_m$ be a basis of V. Define a function $\varphi\colon S_m \to \mathrm{GL}(V)$ by $\sigma \mapsto T_\sigma$, where $T_\sigma : v_i \mapsto v_{\sigma(i)}$ for all i. It is easy to see that φ is an injective homomorphism. By Cayley's

theorem, G can be imbedded in S_G; hence, G can be imbedded in $\mathrm{GL}(m, \mathbb{F}_p)$, where $m = |G|$. Now G is contained in some Sylow p-subgroup P of $\mathrm{GL}(m, \mathbb{F}_p)$, for every p-subgroup lies in some Sylow p-subgroup. Since all Sylow p-subgroups are conjugate, there is $a \in \mathrm{GL}(m, \mathbb{F}_p)$ with $P = a\left(\mathrm{UT}(m, \mathbb{F}_p)\right)a^{-1}$. Therefore,

$$G \cong a^{-1}Ga \leq a^{-1}Pa \leq \mathrm{UT}(m, \mathbb{F}_p). \quad \bullet$$

A natural question is to find the Sylow subgroups of symmetric groups. This can be done, and the answer is in terms of a construction called *wreath product* (see Rotman, *An Introduction to the Theory of Groups*, page 176).

EXERCISES

5.17 How many Sylow 2-subgroups does S_4 have?

5.18 Give an example of a finite group G having Sylow p-subgroups (for some prime p) P, Q and R such that $P \cap Q = \{1\}$ and $P \cap R \neq \{1\}$.
Hint. Consider $S_3 \times S_3$.

5.19 A subgroup H of a group G is called ***characteristic*** if $\varphi(H) \leq H$ for every isomorphism $\varphi \colon G \to G$. A subgroup S of a group G is called ***fully invariant*** if $\varphi(S) \leq S$ for every homomorphism $\varphi \colon G \to G$.

(i) Prove that every fully invariant subgroup is a characteristic subgroup, and that every characteristic subgroup is a normal subgroup.

(ii) Prove that the commutator subgroup, G', is a normal subgroup of G by showing that it is a fully invariant subgroup.

(iii) Give an example of a group G having a normal subgroup H that is not a characteristic subgroup.

(iv) Prove that $Z(G)$, the center of a group G, is a characteristic subgroup (and so $Z(G) \lhd G$), but that it need not be a fully invariant subgroup.
Hint. Let $G = S_3 \times \mathbb{I}_2$.

(v) For any group G, prove that if $H \lhd G$, then $Z(H) \lhd G$.

5.20 If G is an abelian group, prove, for all positive integers m, that mG and $G[m]$ are fully invariant subgroups.

5.21 (**Frattini Argument**). Let K be a normal subgroup of a finite group G. If P is a Sylow p-subgroup of K for some prime p, prove that

$$G = KN_G(P),$$

where $KN_G(P) = \{ab : a \in K \text{ and } b \in N_G(P)\}$.
Hint. If $g \in G$, then gPg^{-1} is a Sylow p-subgroup of K, and so it is conjugate to P in K.

5.22 Prove that $\mathrm{UT}(3, 2) \cong D_8$, and conclude that D_8 is a Sylow 2-subgroup of $\mathrm{GL}(3, 2)$.
Hint. You may use the fact that the only nonabelian groups of order 8 are D_8 and $\mathbf{Q}$.

5.23 (i) Prove that if d is a positive divisor of 24, then S_4 has a subgroup of order d.

(ii) If $d \neq 4$, prove that any two subgroups of S_4 having order d are isomorphic.

5.24 (i) Find a Sylow 3-subgroup of S_6.

Hint. $\{1, 2, 3, 4, 5, 6\} = \{1, 2, 3\} \cup \{4, 5, 6\}$.

(ii) Show that a Sylow 2-subgroup of S_6 is isomorphic to $D_8 \times \mathbb{I}_2$.

Hint. $\{1, 2, 3, 4, 5, 6\} = \{1, 2, 3, 4\} \cup \{5, 6\}$.

5.25 Let Q be a normal p-subgroup of a finite group G. Prove that $Q \leq P$ for every Sylow p-subgroup P of G.

Hint. Use the fact that any other Sylow p-subgroup of G is conjugate to P.

5.26 (i) Let G be a finite group and let P be a Sylow p-subgroup of G. If $H \lhd G$, prove that HP/H is a Sylow p-subgroup of G/H and $H \cap P$ is a Sylow p-subgroup of H.

Hint. Show that $[G/H : HP/H]$ and $[H : H \cap P]$ are prime to p.

(ii) Let P be a Sylow p-subgroup of a finite group G. Give an example of a subgroup H of G with $H \cap P$ not a Sylow p-subgroup of H.

Hint. Choose a subgroup H of S_4 with $H \cong S_3$, and find a Sylow 3-subgroup P of S_4 with $H \cap P = \{1\}$.

5.27 Prove that a Sylow 2-subgroup of A_5 has exactly five conjugates.

5.28 Prove that there are no simple groups of order 96, 120, 300, 312, or 1000.

Hint. Some of these are not tricky.

5.29 Let G be a group of order 90.

(i) If a Sylow 5-subgroup P of G is not normal, prove that it has six conjugates.

Hint. If P has 18 conjugates, there are 72 elements in G of order 5. Show that G has more than 18 other elements.

(ii) Prove that G is not simple.

Hint. Use Exercises 2.95(ii) and 2.96(ii) on page 114.

5.30 Prove that there is no simple group of order 120.

5.31 Prove that there is no simple group of order 150.

5.32 If H is a proper subgroup of a finite group G, prove that G is not the union of all the conjugates of H: that is, $G \neq \bigcup_{x \in G} xHx^{-1}$.

5.3 THE JORDAN–HÖLDER THEOREM

Galois introduced groups to investigate polynomials in $k[x]$, where k is a field of characteristic 0, and he saw that such a polynomial is solvable by radicals if and only if its Galois group is a solvable group. Solvable groups are an interesting family of groups in their own right, and we now examine them a bit more.

Recall that a *normal series* of a group G is a finite sequence of subgroups, $G = G_0, G_1, G_2, \ldots, G_n = \{1\}$, with

$$G = G_0 \geq G_1 \geq G_2 \geq \cdots \geq G_n = \{1\}$$

and $G_{i+1} \lhd G_i$ for all i. The *factor groups* of the series are the groups G_0/G_1, $G_1/G_2, \ldots,$ G_{n-1}/G_n, the *length* of the series is the number of strict inclusions (equivalently, the length

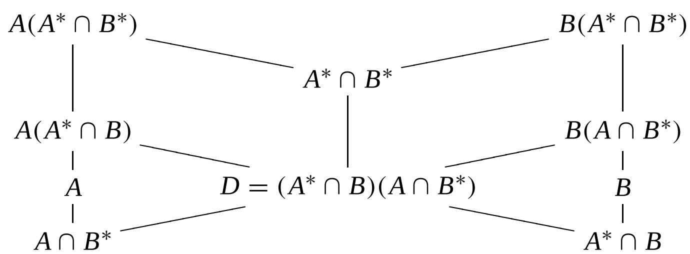
is the number of nontrivial factor groups), and G is *solvable* if it has a normal series whose factor groups are cyclic of prime order.

We begin with a technical result that generalizes the second isomorphism theorem, for we will want to compare different normal series of a group.

Lemma 5.49 (Zassenhaus Lemma). *Given four subgroups $A \lhd A^*$ and $B \lhd B^*$ of a group G, then $A(A^* \cap B) \lhd A(A^* \cap B^*)$, $B(B^* \cap A) \lhd B(B^* \cap A^*)$, and there is an isomorphism*

$$\frac{A(A^* \cap B^*)}{A(A^* \cap B)} \cong \frac{B(B^* \cap A^*)}{B(B^* \cap A)}.$$

Remark. The Zassenhaus lemma is sometimes called the *butterfly lemma* because of the following picture. I confess that I have never liked this picture; it doesn't remind me of a butterfly, and it doesn't help me understand or remember the proof.



The isomorphism is symmetric in the sense that the right side is obtained from the left by interchanging the symbols A and B. ◀

Proof. We claim that $(A \cap B^*) \lhd (A^* \cap B^*)$: that is, if $c \in A \cap B^*$ and $x \in A^* \cap B^*$, then $xcx^{-1} \in A \cap B^*$. Now $xcx^{-1} \in A$ because $c \in A$, $x \in A^*$, and $A \lhd A^*$; but also $xcx^{-1} \in B^*$, because $c, x \in B^*$. Hence, $(A \cap B^*) \lhd (A^* \cap B^*)$; similarly, $(A^* \cap B) \lhd (A^* \cap B^*)$. Therefore, the subset D, defined by $D = (A \cap B^*)(A^* \cap B)$, is a normal subgroup of $A^* \cap B^*$, because it is generated by two normal subgroups.

Using the symmetry in the remark, it suffices to show that there is an isomorphism

$$\frac{A(A^* \cap B^*)}{A(A^* \cap B)} \to \frac{A^* \cap B^*}{D}.$$

Define $\varphi : A(A^* \cap B^*) \to (A^* \cap B^*)/D$ by $\varphi : ax \mapsto xD$, where $a \in A$ and $x \in A^* \cap B^*$. Now φ is well-defined: if $ax = a'x'$, where $a' \in A$ and $x' \in A^* \cap B^*$, then $(a')^{-1}a = x'x^{-1} \in A \cap (A^* \cap B^*) = A \cap B^* \le D$; also, φ is a homomorphism: $axa'x' = a''xx'$, where $a'' = a(xa'x^{-1}) \in A$ (because $A \lhd A^*$), and so $\varphi(axa'x') = \varphi(a''xx') = xx'D = \varphi(ax)\varphi(a'x')$. It is routine to check that φ is surjective and that $\ker \varphi = A(A^* \cap B)$. The first isomorphism theorem completes the proof. •

The reader should check that the Zassenhaus lemma implies the second isomorphism theorem: If S and T are subgroups of a group G with $T \lhd G$, then $TS/T \cong S/(S \cap T)$; set $A^* = G$, $A = T$, $B^* = S$, and $B = S \cap T$.

Definition. A ***composition series*** is a normal series all of whose nontrivial factor groups are simple. The nontrivial factor groups of a composition series are called ***composition factors*** of G.

A group need not have a composition series; for example, the abelian group $\mathbb{Z}$ has no composition series. However, every finite group does have a composition series.

Proposition 5.50. *Every finite group G has a composition series.*

Proof. If the proposition is false, let G be a least criminal; that is, G is a finite group of smallest order that does not have a composition series. Now G is not simple, otherwise $G > \{1\}$ is a composition series. Hence, G has a proper normal subgroup H; we may assume that H is a maximal normal subgroup, so that G/H is a simple group. But $|H| < |G|$, so that H does have a composition series: say, $H = H_0 > H_1 > \cdots > \{1\}$, and $G > H_0 > H_1 > \cdots > \{1\}$ is a composition series for G, a contradiction. •

A group G is solvable if it has a normal series with factor groups cyclic of prime order. As cyclic groups of prime order are simple groups, a normal series as in the definition of solvable group is a composition series, and so composition factors of G are cyclic groups of prime order.

Here are two composition series of $G = \langle a \rangle$, a cyclic group of order 30 (note that normality of subgroups is automatic because G is abelian). The first is

$$G = \langle a \rangle \geq \langle a^2 \rangle \geq \langle a^{10} \rangle \geq \{1\};$$

the factor groups of this series are $\langle a \rangle/\langle a^2 \rangle \cong \mathbb{I}_2$, $\langle a^2 \rangle/\langle a^{10} \rangle \cong \mathbb{I}_5$, and $\langle a^{10} \rangle/\{1\} \cong \langle a^{10} \rangle \cong \mathbb{I}_3$. Another normal series is

$$G = \langle a \rangle \geq \langle a^5 \rangle \geq \langle a^{15} \rangle \geq \{1\};$$

the factor groups of this series are $\langle a \rangle/\langle a^5 \rangle \cong \mathbb{I}_5$, $\langle a^5 \rangle/\langle a^{15} \rangle \cong \mathbb{I}_3$, and $\langle a^{15} \rangle/\{1\} \cong \langle a^{15} \rangle \cong \mathbb{I}_2$. Notice that the same factor groups arise, although the order in which they arise is different. We will see that this phenomenon always occurs: Different composition series of the same group have the same factor groups. This is the *Jordan–Hölder theorem*, and the next definition makes its statement more precise.

Definition. Two normal series of a group G are ***equivalent*** if there is a bijection between the sets of nontrivial factor groups of each so that corresponding factor groups are isomorphic.

The Jordan–Hölder theorem says that any two composition series of a group are equivalent. It will be more efficient to prove a more general theorem, due to Schreier.

Definition. A ***refinement*** of a normal series is a normal series $G = N_0, N_1, \ldots, N_k = \{1\}$ having the original series as a subsequence.

In other words, a refinement of a normal series is a new normal series obtained from the original by inserting more subgroups.

Notice that a composition series admits only insignificant refinements; one can merely repeat terms (if G_i/G_{i+1} is simple, then it has no proper nontrivial normal subgroups and, hence, there is no intermediate subgroup L with $G_i > L > G_{i+1}$ and $L \lhd G_i$). Therefore, any refinement of a composition series is equivalent to the original composition series.

Theorem 5.51 (Schreier Refinement Theorem). *Any two normal series*

$$G = G_0 \geq G_1 \geq \cdots \geq G_n = \{1\}$$

and

$$G = N_0 \geq N_1 \geq \cdots \geq N_k = \{1\}$$

of a group G have equivalent refinements.

Proof. We insert a copy of the second series between each pair of adjacent terms in the first series. In more detail, for each $i \geq 0$, define

$$G_{ij} = G_{i+1}(G_i \cap N_j)$$

(this is a subgroup because $G_{i+1} \lhd G_i$). Note that

$$G_{i0} = G_{i+1}(G_i \cap N_0) = G_{i+1}G_i = G_i,$$

because $N_0 = G$, and that

$$G_{ik} = G_{i+1}(G_i \cap N_k) = G_{i+1},$$

because $N_k = \{1\}$. Therefore, the series of G_{ij} is a subsequence of the series of G_i:

$$\cdots \geq G_i = G_{i0} \geq G_{i1} \geq G_{i2} \geq \cdots \geq G_{ik} = G_{i+1} \geq \cdots .$$

Similarly, there is a subsequence of the second series arising from subgroups

$$N_{pq} = N_{p+1}(N_p \cap G_q).$$

Both subsequences have nk terms. For each i, j, the Zassenhaus lemma, for the four subgroups $G_{i+1} \lhd G_i$ and $N_{j+1} \lhd N_j$, says both subsequences are normal series, hence are refinements, and there is an isomorphism

$$\frac{G_{i+1}(G_i \cap N_j)}{G_{i+1}(G_i \cap N_{j+1})} \cong \frac{N_{j+1}(N_j \cap G_i)}{N_{j+1}(N_j \cap G_{i+1})};$$

that is,

$$G_{i,j}/G_{i,j+1} \cong N_{j,i}/N_{j,i+1}.$$

The association $G_{i,j}/G_{i,j+1} \mapsto N_{j,i}/N_{j,i+1}$ is a bijection showing that the two refinements are equivalent. •

Theorem 5.52 (Jordan–Hölder[4] Theorem). *Any two composition series of a group G are equivalent. In particular, the length of a composition series, if one exists, is an invariant of G.*

Proof. As we remarked earlier, any refinement of a composition series is equivalent to the original composition series. It now follows from Schreier's theorem that any two composition series are equivalent. •

Here is a new proof of the fundamental theorem of arithmetic.

Corollary 5.53. *Every integer $n \geq 2$ has a factorization into primes, and the prime factors are uniquely determined by n.*

Proof. Since the group $\mathbb{I}_n$ is finite, it has a composition series; let $S_1, \ldots, S_t$ be the factor groups. Now an abelian group is simple if and only if it is of prime order, by Proposition 2.107; since $n = |\mathbb{I}_n|$ is the product of the orders of the factor groups (see Exercise 5.36 on page 287), we have proved that n is a product of primes. Moreover, the Jordan–Hölder theorem gives the uniqueness of the (prime) orders of the factor groups. •

Example 5.54.

(i) Nonisomorphic groups can have the same composition factors. For example, both $\mathbb{I}_4$ and $\mathbf{V}$ have composition series whose factor groups are $\mathbb{I}_2, \mathbb{I}_2$.

(ii) Let $G = \mathrm{GL}(2, \mathbb{F}_4)$ be the general linear group of all 2×2 nonsingular matrices with entries in the field $\mathbb{F}_4$ with four elements. Now $\det\colon G \to (\mathbb{F}_4)^\times$, where $(\mathbb{F}_4)^\times \cong \mathbb{I}_3$ is the multiplicative group of nonzero elements of $\mathbb{F}_4$. Since $\ker \det = \mathrm{SL}(2, \mathbb{F}_4)$, the special linear group consisting of those matrices of determinant 1, there is a normal series

$$G = \mathrm{GL}(2, \mathbb{F}_4) \geq \mathrm{SL}(2, \mathbb{F}_4) \geq \{1\}.$$

The factor groups of this normal series are $\mathbb{I}_3$ and $\mathrm{SL}(2, \mathbb{F}_4)$. It is true that $\mathrm{SL}(2, \mathbb{F}_4)$ is a nonabelian simple group [in fact, Corollary 5.68 says that $\mathrm{SL}(2, \mathbb{F}_4) \cong A_5$], and so this series is a composition series. We cannot yet conclude that G is not solvable, for the definition of solvability requires that there be some composition series, not necessarily this one, having factor groups of prime order. However, the Jordan–Hölder theorem says that if one composition series of G has all its factor groups of prime order, then so does every other composition series. We may now conclude that $\mathrm{GL}(2, \mathbb{F}_4)$ is not a solvable group. ◂

We now discuss the importance of the Jordan–Hölder theorem in group theory.

Definition. If G is a group and $K \lhd G$, then G is called an ***extension*** of K by G/K.

[4] In 1868, C. Jordan proved that the orders of the factor groups of a composition series depend only on G and not upon the composition series; in 1889, O. Hölder proved that the factor groups themselves, to isomorphism, do not depend upon the composition series.

With this terminology, Exercise 2.75 on page 95 says that an extension of one p-group by another p-group is itself a p-group, and Proposition 4.24 says that any extension of one solvable group by another is itself a solvable group.

The study of extensions involves the inverse question: How much of G can be recovered from a normal subgroup K and the quotient $Q = G/K$? For example, we do know that if K and Q are finite, then $|G| = |K||Q|$.

Example 5.55.
(i) The direct product $K \times Q$ is an extension of K by Q (and $K \times Q$ is an extension of Q by K).

(ii) Both S_3 and $\mathbb{I}_6$ are extensions of $\mathbb{I}_3$ by $\mathbb{I}_2$. On the other hand, $\mathbb{I}_6$ is an extension of $\mathbb{I}_2$ by $\mathbb{I}_3$, but S_3 is not, for S_3 contains no normal subgroup of order 2. ◂

We have just seen, for any given pair of groups K and Q, that an extension of K by Q always exists (the direct product), but there may be nonisomorphic such extensions. Hence, if we view an extension of K by Q as a "product" of K and Q, then this product is not single-valued. The ***extension problem*** is to classify all possible extensions of a given pair of groups K and Q.

Suppose that a group G has a normal series

$$G = K_0 \geq K_1 \geq K_2 \geq \cdots \geq K_{n-1} \geq K_n = \{1\}$$

with factor groups $Q_1, \ldots, Q_n$, where

$$Q_i = K_{i-1}/K_i$$

for all $i \geq 1$. Now $K_n = \{1\}$, so that $K_{n-1} = Q_n$, but something more interesting occurs next: $K_{n-2}/K_{n-1} = Q_{n-1}$, so that K_{n-2} is an extension of K_{n-1} by Q_{n-1}. If we could solve the extension problem, then we could recapture K_{n-2} from K_{n-1} and Q_{n-1}—that is, from Q_n and Q_{n-1}. Next, observe that $K_{n-3}/K_{n-2} = Q_{n-2}$, so that K_{n-3} is an extension of K_{n-2} by Q_{n-2}. If we could solve the extension problem, then we could recapture K_{n-3} from K_{n-2} and Q_{n-2}; that is, we could recapture K_{n-3} from Q_n, Q_{n-1}, and Q_{n-2}. Climbing up the composition series in this way, we end with $G = K_0$ being recaptured from $Q_n, Q_{n-1}, \ldots, Q_1$. Thus, G is a "product" of the factor groups. If the normal series is a composition series, then the Jordan–Hölder theorem says that the factors in this product (that is, the composition factors of G) are uniquely determined by G. Therefore, we could survey all finite groups if we knew the finite simple groups and if we could solve the extension problem. Now all the finite simple groups were classified in the 1980s; this theorem, one of the deepest theorems in mathematics, gives a complete list of all the finite simple groups, along with interesting properties of them. In a sense, the extension problem has also been solved. In Chapter 10, we will give a solution to the extension problem, due to Schreier, which describes all possible multiplication tables for extensions; this study leads to *cohomology of groups* and the *Schur–Zassenhaus theorem*. On the other hand, the extension problem is unsolved in that no one knows a way, given K and Q, to compute the exact number of nonisomorphic extensions of K by Q.

We now pass from general groups (whose composition factors are arbitrary simple groups) to solvable groups (whose composition factors are cyclic groups of prime order; cyclic groups of prime order are simple in every sense of the word). Even though solvable groups arose in determining those polynomials that are solvable by radicals, there are purely group-theoretic theorems about solvable groups making no direct reference to Galois theory and polynomials. For example, a theorem of P. Hall generalizes the Sylow theorems as follows: If G is a solvable group of order ab, where a and b are relatively prime, then G contains a subgroup of order a; moreover, any two such subgroups are conjugate. A theorem of W. Burnside says that if $|G| = p^m q^n$, where p and q are prime, then G is solvable. The remarkable *Feit–Thompson theorem* states that every group of odd order must be solvable.

Solvability of a group is preserved by standard group-theoretic constructions. For example, we have seen, in Proposition 4.21, that every quotient G/N of a solvable group G is itself a solvable group, while Proposition 4.22 shows that every subgroup of a solvable group is itself solvable. Proposition 4.24 shows that an extension of one solvable group by another is itself solvable: If $H \lhd G$ and both H and G/H are solvable, then G is solvable, and Corollary 4.25 shows that a direct product of solvable groups is itself solvable.

Proposition 5.56. *Every finite p-group G is solvable.*

Proof. If G is abelian, then G is solvable. Otherwise, its center, $Z(G)$, is a proper nontrivial normal abelian subgroup, by Theorem 2.103. Now $Z(G)$ is solvable, because it is abelian, and $G/Z(G)$ is solvable, by induction on $|G|$, and so G is solvable, by Proposition 4.24. •

It follows, of course, that a direct product of finite p-groups is solvable.

Definition. If G is a group and $x, y \in G$, then their ***commutator*** $[x, y]$ is the element

$$[x, y] = xyx^{-1}y^{-1}.$$

If X and Y are subgroups of a group G, then $[X, Y]$ is defined by

$$[X, Y] = \langle [x, y] : x \in X \text{ and } y \in Y \rangle.$$

In particular, the ***commutator subgroup*** G' of a group G is

$$G' = [G, G],$$

the subgroup generated by all the commutators.[5]

It is clear that two elements x and y in a group G commute if and only if their commutator $[x, y]$ is 1. The next proposition generalizes this observation.

[5] The subset consisting of all the commutators need not be closed under products, and so the *set* of all commutators may not be a subgroup. The smallest group in which a product of two commutators is not a commutator has order 96. Also, see Carmichael's exercise on page 297.

Proposition 5.57. *Let G be a group.*

(i) *The commutator subgroup G' is a normal subgroup of G, and G/G' is abelian.*

(ii) *If $H \lhd G$ and G/H is abelian, then $G' \le H$.*

Proof. (i) The inverse of a commutator $xyx^{-1}y^{-1}$ is itself a commutator: $[x, y]^{-1} = yxy^{-1}x^{-1} = [y, x]$. Therefore, each element of G' is a product of commutators. But any conjugate of a commutator (and hence, a product of commutators) is another such:

$$\begin{aligned} a[x, y]a^{-1} &= a(xyx^{-1}y^{-1})a^{-1} \\ &= axa^{-1}aya^{-1}ax^{-1}a^{-1}ay^{-1}a^{-1} \\ &= [axa^{-1}, aya^{-1}]. \end{aligned}$$

Therefore, $G' \lhd G$. (Alternatively, G' is a fully invariant subgroup of G, for if $\varphi \colon G \to G$ is a homomorphism, then $\varphi([x, y]) = [\varphi(x), \varphi(y)] \in G'$.)

If $aG', bG' \in G/G'$, then

$$aG'bG'(aG')^{-1}(bG')^{-1} = aba^{-1}b^{-1}G' = [a, b]G' = G',$$

and so G/G' is abelian.

(ii) Suppose that $H \lhd G$ and G/H is abelian. If $a, b \in G$, then $aHbH = bHaH$; that is, $abH = baH$, and so $b^{-1}a^{-1}ba \in H$. As every commutator has the form $b^{-1}a^{-1}ba$, we have $G' \le H$. •

Example 5.58.

(i) A group G is abelian if and only if $G' = \{1\}$.

(ii) If G is a simple group, then $G' = \{1\}$ or $G' = G$, for G' is a normal subgroup. The first case occurs when G has prime order; the second case occurs otherwise. In particular, $(A_n)' = A_n$ for all $n \ge 5$.

(iii) We show that $(S_n)' = A_n$ for all $n \ge 5$. Since $S_n/A_n \cong \mathbb{I}_2$ is abelian, Proposition 5.57 shows that $(S_n)' \le A_n$. For the reverse inclusion, note that $(S_n)' \cap A_n \lhd A_n$, so that the simplicity of A_n gives this intersection trivial or A_n. Clearly, $(S_n)' \cap A_n \ne \{(1)\}$, and so $A_n \le (S_n)'$. ◄

Let us iterate the formation of the commutator subgroup.

Definition. The ***derived series*** of G is

$$G = G^{(0)} \ge G^{(1)} \ge G^{(2)} \ge \cdots \ge G^{(i)} \ge G^{(i+1)} \ge \cdots,$$

where $G^{(0)} = G$, $G^{(1)} = G'$, and, more generally, $G^{(i+1)} = (G^{(i)})' = [G^{(i)}, G^{(i)}]$ for all $i \ge 0$.

It is easy to prove by induction on $i \geq 0$, that $G^{(i)}$ is fully invariant, which implies that $G^{(i)} \lhd G$; it follows that $G^{(i+1)} \lhd G^{(i)}$, and so the derived series is a normal series. The derived series can be used to give a characterization of solvability: G is ***solvable*** if and only if the derived series reaches $\{1\}$.

Proposition 5.59.

(i) *A finite group G is solvable if and only if it has a normal series with abelian factor groups.*

(ii) *A finite group G is solvable if and only if there is some n with*

$$G^{(n)} = \{1\}.$$

Proof. (i) If G is solvable, then it has a normal series whose factor groups G_i/G_{i+1} are all cyclic of prime order, hence are abelian.

Conversely, if G has a normal series with abelian factor groups, then the factor groups of any refinement are also abelian. In particular, the factor groups of a composition series of G, which exists because G is finite, are abelian simple groups; hence, they are cyclic of prime order, and so G is solvable.

(ii) Assume that G is solvable, so there is a normal series

$$G \geq G_1 \geq G_2 \geq \cdots \geq G_n = \{1\}$$

whose factor groups G_i/G_{i+1} are abelian. We show, by induction on $i \geq 0$, that $G^{(i)} \leq G_i$. Since $G^{(0)} = G = G_0$, the base step is obviously true. For the inductive step, since G_i/G_{i+1} is abelian, Proposition 5.57 gives $(G_i)' \leq G_{i+1}$. On the other hand, the inductive hypothesis gives $G^{(i)} \leq G_i$, which implies that

$$G^{(i+1)} = (G^{(i)})' \leq (G_i)' \leq G_{i+1}.$$

In particular, $G^{(n)} \leq G_n = \{1\}$, which is what we wished to show.

Conversely, if $G^{(n)} = \{1\}$, then the derived series is a normal series (a normal series must end with $\{1\}$) with abelian factor groups, and so part (i) gives G solvable. •

For example, the derived series of $G = S_4$ is easily seen to be

$$S_4 > A_4 > \mathbf{V} > \{(1)\}.$$

Our earlier definition of solvability applies only to finite groups, whereas the characterization in the proposition makes sense for all groups, possibly infinite. Nowadays, most authors define a group to be solvable if its derived series reaches $\{1\}$ after a finite number of steps; with this new definition, every abelian group is solvable, whereas it is easy to see that abelian groups are solvable in the sense of the original definition if and only if they are finite. In Exercise 5.38 on page 287, the reader will be asked to prove, using the criterion in Proposition 5.59, that subgroups, quotient groups, and extensions of solvable groups are also solvable (in the new, generalized, sense).

There are other interesting classes of groups defined in terms of normal series. One of the most interesting such consists of *nilpotent* groups.

Definition. The ***descending central series*** of a group G is

$$G = \gamma_1(G) \geq \gamma_2(G) \geq \cdots,$$

where $\gamma_{i+1}(G) = [\gamma_i(G), G]$. A group G is called ***nilpotent*** if the lower central series reaches $\{1\}$; that is, if $\gamma_n(G) = \{1\}$ for some n.

Note that $\gamma_2(G) = G'$, but the derived series and the lower central series may differ afterward; for example, $\gamma_3(G) = [G', G] \geq G^{(2)}$, with strict inequality possible.

Finite nilpotent groups can be characterized by Proposition 5.39: they are the groups that are direct products of their Sylow subgroups, and so one regards finite nilpotent groups as generalized p-groups. Examples of nilpotent groups are $\mathrm{UT}(n, \mathbb{F}_q)$, $\mathrm{UT}(n, \mathbb{Z})$ (unitriangular groups over $\mathbb{Z}$), the Frattini subgroup $\Phi(G)$ (defined in Exercise 5.46 on page 288) of a finite group G, and certain automorphism groups arising from a normal series of a group. We can prove results, for infinite nilpotent groups as well as for finite ones, such as those in Exercise 5.47 on page 288: Every subgroup and every quotient of a finite nilpotent group G is again nilpotent; if $G/Z(G)$ is nilpotent, then so is G; every normal subgroup H intersects $Z(G)$ nontrivially.

EXERCISES

5.33 Let p be a prime and let G be a nonabelian group of order p^3. Prove that $Z(G) = G'$.
Hint. Show first that both subgroups have order p.

5.34 Prove that if H is a subgroup of a group G and $G' \leq H$, then $H \lhd G$.
Hint. Use the correspondence theorem.

5.35 (i) Prove that $(S_n)' = A_n$ for $n = 2, 3, 4$ [see Example 5.58(iii) for $n \geq 5$].
(ii) Prove that $(\mathrm{GL}(n, k))' \leq \mathrm{SL}(n, k)$. (The reverse inclusion is also true; see Exercise 5.56 on page 296 for the case $n = 2$.)

5.36 If G is a finite group and

$$G = G_0 \geq G_1 \geq \cdots \geq G_n = \{1\}$$

is a normal series, prove that the order of G is the product of the orders of the factor groups:

$$|G| = \prod_{i=0}^{n-1} |G_i/G_{i+1}|.$$

5.37 Prove that any two finite solvable groups of the same order have the same composition factors.

5.38 Let G be an arbitrary, possibly infinite group.
(i) Prove that if $H \leq G$, then $H^{(i)} \leq G^{(i)}$ for all i. Conclude, using Proposition 5.59, that every subgroup of a solvable group is solvable.
(ii) Prove that if $f : G \to K$ is a surjective homomorphism, then

$$f(G^{(i)}) = K^{(i)}$$

for all i. Conclude, using Proposition 5.59, that every quotient of a solvable group is also solvable.

(iii) For every group G, prove, by double induction, that

$$G^{(m+n)} = (G^{(m)})^{(n)}.$$

(iv) Prove, using Proposition 5.59, that if $H \lhd G$ and both H and G/H are solvable, then G is solvable.

5.39 Let p and q be primes.

(i) Prove that every group of order pq is solvable.

Hint. If $p = q$, then G is abelian. If $p < q$, then a divisor r of pq for which $r \equiv 1 \bmod q$ must equal 1.

(ii) Prove that every group G of order p^2q is solvable.

Hint. If G is not simple, use Proposition 4.24. If $p > q$, then $r \equiv 1 \bmod p$ forces $r = 1$. If $p < q$, then $r = p^2$ and there are more than p^2q elements in G.

5.40 Show that the Feit–Thompson theorem—"Every finite group of odd order is solvable," is equivalent to "Every nonabelian finite simple group has even order."
Hint. For sufficiency, choose a "least criminal": a nonsolvable group G of smallest odd order. By hypothesis, G is not simple, and so it has a proper nontrivial normal subgroup.

5.41 (i) Prove that the infinite cyclic group $\mathbb{Z}$ does not have a composition series.

(ii) Prove that an abelian group G has a composition series if and only if G is finite.

5.42 Prove that if G is a finite group and $H \lhd G$, then there is a composition series of G one of whose terms is H.
Hint. Use Schreier's theorem.

5.43 (i) Prove that if S and T are solvable subgroups of a group G and $S \lhd G$, then ST is a solvable subgroup of G.

Hint. The subgroup ST is a homomorphic image of $S \times T$.

(ii) If G is a finite group, define $\mathcal{S}(G)$ to be the subgroup of G generated by all normal solvable subgroups of G. Prove that $\mathcal{S}(G)$ is the unique maximal normal solvable subgroup of G and that $G/\mathcal{S}(G)$ has no nontrivial normal solvable subgroups.

5.44 (i) Prove that the dihedral groups D_{2n} are solvable.

(ii) Give a composition series for D_{2n}.

5.45 (***Rosset***). Let G be a group containing elements x and y such that the orders of x, y, and xy are pairwise relatively prime; prove that G is not solvable.

5.46 (i) If G is a finite group, then its ***Frattini subgroup***, denoted by $\Phi(G)$, is defined to be the intersection of all the maximal subgroups of G. Prove that $\Phi(G)$ is a characteristic subgroup, and hence it is a normal subgroup of G.

(ii) Prove that if p is a prime and G is a finite abelian p-group, then $\Phi(G) = pG$. The ***Burnside basis theorem*** says that if G is any (not necessarily abelian) finite p-group, then $G/\Phi(G)$ is a vector space over $\mathbb{F}_p$, and its dimension is the minimum number of generators of G (see Rotman, *An Introduction to the Theory of Groups,* page 124).

5.47 (i) If G is a nilpotent group, prove that its center $Z(G) \neq \{1\}$.

(ii) If G is a group with $G/Z(G)$ nilpotent, prove that G is nilpotent.

(iii) If G is a nilpotent group, prove that every subgroup and every quotient group of G is also nilpotent.

(iv) Let G be a group and let $H \lhd G$. Give an example in which both H and G/H are nilpotent and yet G is not nilpotent.

(v) If G is a finite p-group and if $H \lhd G$, prove that $H \cap Z(G) \neq \{1\}$. (The generalization of this result to finite nilpotent groups is true.)

5.48 Let $\mathfrak{A}$ denote the class of all abelian groups, $\mathfrak{N}$ the class of all nilpotent groups, and $\mathfrak{S}$ the class of all solvable groups.

(i) Prove that $\mathfrak{A} \subseteq \mathfrak{N} \subseteq \mathfrak{S}$.

(ii) Show that each of the inclusions in part (i) is strict; that is, there is a nilpotent group that is not abelian, and there is a solvable group that is not nilpotent.

5.49 If G is a group and $g, x \in G$, write $g^x = xgx^{-1}$.

(i) Prove, for all $x, y, z \in G$, that $[x, yz] = [x, y][x, z]^y$ and $[xy, z] = [y, z]^x[x, z]$.

(ii) (***Jacobi Identity***) If $x, y, z \in G$ are elements in a group G, define

$$[x, y, z] = [x, [y, z]].$$

Prove that

$$[x, y^{-1}, z]^y[y, z^{-1}, x]^z[z, x^{-1}, y]^x = 1.$$

5.50 If H, K, L are subgroups of a group G, define

$$[H, K, L] = \langle\{[h, k, \ell] : h \in H, k \in K, \ell \in L\}\rangle.$$

(i) Prove that if $[H, K, L] = \{1\} = [K, L, H]$, then $[L, H, K] = \{1\}$.

(ii) (***Three subgroups lemma***) If $N \lhd G$ and $[H, K, L][K, L, H] \leq N$, prove that

$$[L, H, K] \leq N.$$

(iii) Prove that if G is a group with $G = G'$, then $G/Z(G)$ is centerless.

Hint. If $\pi : G \to G/Z(G)$ is the natural map, define $\zeta^2(G) = \pi^{-1}(Z(G/Z(G)))$. Use the three subgroups lemma with $L = \zeta^2(G)$ and $H = K = G$.

(iv) Prove, for all i, j, that $[\gamma_i(G), \gamma_j(G)] \leq \gamma_{i+j}(G)$.

5.4 Projective Unimodular Groups

The Jordan-Hölder theorem associates a family of simple groups to every finite group, and it can be used to reduce many problems about finite groups to problems about finite simple groups. This empirical fact says that a knowledge of simple groups is very useful. The only simple groups we have seen so far are cyclic groups of prime order and the alternating groups A_n for $n \geq 5$. We will now show that certain finite groups of matrices are simple, and we begin by considering some matrices that will play the same role for 2×2 linear groups as the 3-cycles played for the alternating groups.

Definition. A ***transvection***[6] over a field k is a matrix of the form

$$B_{12}(r) = \begin{bmatrix} 1 & r \\ 0 & 1 \end{bmatrix} \quad \text{or} \quad B_{21}(r) = \begin{bmatrix} 1 & 0 \\ r & 1 \end{bmatrix},$$

where $r \in k$ and $r \neq 0$.

Let A be a 2×2 matrix. It is easy to see that $B_{12}(r)A$ is the matrix obtained from A by replacing Row(1) by Row(1) + rRow(2), and that $B_{21}(r)A$ is the matrix obtained from A by replacing Row(2) by Row(2) + rRow(1).

Lemma 5.60. *If k is a field and $A \in \mathrm{GL}(2, k)$, then*

$$A = UD,$$

where U is a product of transvections and $D = \mathrm{diag}\{1, d\} = \begin{bmatrix} 1 & 0 \\ 0 & d \end{bmatrix}$, where $d = \det(A)$.

Proof. Let

$$A = \begin{bmatrix} p & q \\ r & s \end{bmatrix}.$$

We may assume that $r \neq 0$; otherwise, $p \neq 0$ (because A is nonsingular), and replacing Row(2) by Row(2)+Row(1) puts p in the 21 position. Next, replace Row(1) by Row(1)+ $r^{-1}(1 - p)$Row(2), so that 1 is in the upper left corner. Now continue multiplying by transvections:

$$\begin{bmatrix} 1 & x \\ r & s \end{bmatrix} \to \begin{bmatrix} 1 & x \\ 0 & y \end{bmatrix} \to \begin{bmatrix} 1 & 0 \\ 0 & d \end{bmatrix}.$$

Thus, $WA = D$, where W is a product of transvections and $D = \mathrm{diag}\{1, d\}$. Since transvections have determinant 1, we have $\det(W) = 1$, and so

$$\det(A) = \det(D) = d.$$

As the inverse of a transvection is also a transvection, we have $A = W^{-1}D$, which is the factorization we seek. •

Recall that $\mathrm{SL}(2, k)$ is the subgroup of $\mathrm{GL}(2, k)$ consisting of all matrices of determinant 1.[7] If k is a finite field, then $k \cong \mathbb{F}_q$, where $q = p^n$ and p is a prime; we may denote $\mathrm{GL}(2, \mathbb{F}_q)$ by $\mathrm{GL}(2, q)$ and, similarly, we may denote $\mathrm{SL}(2, \mathbb{F}_q)$ by $\mathrm{SL}(2, q)$.

---

[6] Most group theorists define a 2×2 transvection as a matrix that is *similar* to $B_{12}(r)$ or $B_{21}(r)$ [that is, a conjugate of $B_{12}(r)$ or $B_{21}(r)$ in $\mathrm{GL}(2, k)$]. The word *transvection* is a synonym for *transporting*, and its usage in this context is probably due to E. Artin, who gives the following definition in his book *Geometric Algebra*: "An element $\tau \in \mathrm{GL}(V)$, where V is an n-dimensional vector space, is called a ***transvection*** if it keeps every vector of some hyperplane H fixed and moves any vector $x \in V$ by some vector of H; that is, $\tau(x) - x \in H$." In our case, $B_{12}(r)$ fixes the "x-axis" and $B_{21}(r)$ fixes the "y-axis."

[7] GL abbreviates *general linear* and SL abbreviates *special linear*.

Proposition 5.61.

(i) *If k is a field, then* $\mathrm{SL}(2, k)$ *is generated by transvections.*

(ii) *If k is a field, then* $\mathrm{GL}(2, k)/\mathrm{SL}(2, k) \cong k^\times$*, where $k^\times$ is the multiplicative group of nonzero elements of k.*

(iii) *If $k = \mathbb{F}_q$, then*

$$|\mathrm{SL}(2, \mathbb{F}_q)| = (q + 1)q(q - 1).$$

Proof. (i) If $A \in \mathrm{SL}(2, k)$, then Lemma 5.60 gives a factorization $A = UD$, where U is a product of transvections and $D = \mathrm{diag}\{1, d\}$, where $d = \det(A)$. Since $A \in \mathrm{SL}(2, k)$, we have $\det(A) = 1$, and so $A = U$.

(ii) If $a \in k^\times$, then the matrix $\mathrm{diag}\{1, a\}$ has determinant a, hence is nonsingular, and so the map $\det\colon \mathrm{GL}(2, k) \to k^\times$ is surjective. The definition of $\mathrm{SL}(2, k)$ shows that it is the kernel of det, and so the first isomorphism theorem gives the result.

(iii) If H is a normal subgroup of a finite group G, then Lagrange's theorem gives $|H| = |G|/|G/H|$. In particular,

$$|\mathrm{SL}(2, \mathbb{F}_q)| = |\,\mathrm{GL}(2, \mathbb{F}_q)|/|\mathbb{F}_q^\times|.$$

But $|\,\mathrm{GL}(2, \mathbb{F}_q)| = (q^2 - 1)(q^2 - q)$, by Theorem 5.46, and $|\mathbb{F}_q^\times| = q - 1$. Hence, $|\mathrm{SL}(2, \mathbb{F}_q)| = (q + 1)q(q - 1)$. •

We now compute the center of these matrix groups. If V is a two-dimensional vector space over k, then we proved, in Proposition 3.108, that $\mathrm{GL}(2, k) \cong \mathrm{GL}(V)$, the group of all nonsingular linear transformations on V. Moreover, Proposition 3.109(i) identifies the center with the scalar transformations.

Proposition 5.62. *The center of* $\mathrm{SL}(2, k)$*, denoted by* $\mathrm{SZ}(2, k)$*, consists of all scalar matrices* $\begin{bmatrix} a & 0 \\ 0 & a \end{bmatrix}$ *with $a^2 = 1$.*

Remark. Here we see that $\mathrm{SZ} = \mathrm{SL} \cap Z(\mathrm{GL})$, but it is not true in general that if $H \le G$, then $Z(H) = H \cap Z(G)$ (indeed, this equality may not hold even when H is normal). We always have $H \cap Z(G) \le Z(H)$, but the inclusion may be strict. For example, if $G = S_3$ and $H = A_3 \cong \mathbb{I}_3$, then $Z(A_3) = A_3$ while $A_3 \cap Z(S_3) = \{1\}$. ◂

Proof. It is more convenient here to use linear transformations than matrices. Assume that $T \in \mathrm{SL}(2, k)$ is not a scalar transformation. Therefore, there is a nonzero vector $v \in V$ with Tv not a scalar multiple of v. It follows that the list v, Tv is linearly independent and, since $\dim(V) = 2$, that it is a basis of V. Define $S : V \to V$ by $S(v) = v$ and $S(Tv) = v + Tv$. Notice, relative to the basis v, Tv, that S has matrix $B_{12}(1)$, so that $\det(S) = 1$. Now T and S do not commute, for $TS(v) = Tv$ while $ST(v) = v + Tv$. It follows that the center must consist of scalar transformations. In matrix terms, the center consists of scalar matrices $A = \mathrm{diag}\{a, a\}$, and $a^2 = \det(A) = 1$. •

Definition. The ***projective unimodular***[8] ***group*** is the quotient group

$$\mathrm{PSL}(2,k) = \mathrm{SL}(2,k)/\mathrm{SZ}(2,k).$$

Note that if $c^2 = 1$, where c is in a field k, then $c = \pm 1$. If $k = \mathbb{F}_q$, where q is a power of 2, then $\mathbb{F}_q$ has characteristic 2, so that $c^2 = 1$ implies $c = 1$. Therefore, in this case, $\mathrm{SZ}(2,\mathbb{F}_q) = \{I\}$ and so $\mathrm{PSL}(2,\mathbb{F}_{2^n}) = \mathrm{SL}(2,\mathbb{F}_{2^n})$.

Proposition 5.63.

$$|\mathrm{PSL}(2,\mathbb{F}_q)| = \begin{cases} \frac{1}{2}(q+1)q(q-1) & \textit{if } q = p^n \textit{ and } p \textit{ is an odd prime}; \\ (q+1)q(q-1) & \textit{if } q = 2^n. \end{cases}$$

Proof. Proposition 5.61(iii) gives $|\mathrm{PSL}(2,\mathbb{F}_q)| = (q+1)q(q-1)/|\mathrm{SZ}(2,\mathbb{F}_q)|$ and, Proposition 5.62 gives

$$|\mathrm{SZ}(2,\mathbb{F}_q)| = |\{a \in \mathbb{F}_q : a^2 = 1\}|.$$

Now $\mathbb{F}_q^\times$ is a cyclic group of order $q-1$, by Theorem 3.30. If q is odd, then $q-1$ is even, and the cyclic group $\mathbb{F}_q^\times$ has a unique subgroup of order 2; if q is a power of 2, then we noted, just before the statement of this proposition, that $\mathrm{SZ}(2,\mathbb{F}_q) = \{I\}$. Therefore, $|SZ(2,q)| = 2$ if q is a power of an odd prime, and $|SZ(2,q)| = 1$ if q is a power of 2. •

We are now going to prove that the groups $\mathrm{PSL}(2,\mathbb{F}_q)$ are simple for all prime powers $q \geq 4$. As we said earlier, the transvections will play the role of the 3-cycles (see Exercise 2.91 on page 113).

Lemma 5.64. *If H is a normal subgroup of* $\mathrm{SL}(2,\mathbb{F}_q)$ *containing a transvection* $B_{12}(r)$ *or* $B_{21}(r)$*, then* $H = \mathrm{SL}(2,\mathbb{F}_q)$.

Proof. Note first that if

$$U = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix},$$

then $\det(U) = 1$ and $U \in \mathrm{SL}(2,\mathbb{F}_q)$; since H is a normal subgroup, $UB_{12}(r)U^{-1}$ also lies in H. But $UB_{12}(r)U^{-1} = B_{21}(-r)$, from which it follows that H contains a transvection of the form $B_{12}(r)$ if and only if it contains a transvection of the form $B_{21}(-r)$. Since SL is generated by the transvections, it suffices to show that every transvection $B_{12}(r)$ lies in H.

The following conjugate of $B_{12}(r)$ lies in H because H is normal:

$$\begin{bmatrix} \alpha & \beta \\ 0 & \alpha^{-1} \end{bmatrix} \begin{bmatrix} 1 & r \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \alpha^{-1} & -\beta \\ 0 & \alpha \end{bmatrix} = \begin{bmatrix} 1 & r\alpha^2 \\ 0 & 1 \end{bmatrix} = B_{12}(r\alpha^2).$$

Define

$$G = \{0\} \cup \{u \in \mathbb{F}_q : B_{12}(u) \in H\}.$$

[8] A matrix is called *unimodular* if it has determinant 1. The adjective *projective* arises because this group turns out to consist of automorphisms of a projective plane.

We have just shown that $r\alpha^2 \in G$ for all $\alpha \in \mathbb{F}_q$. It is easy to check that G is a subgroup of the additive group of $\mathbb{F}_q$ and, hence, it contains all the elements of the form $u = r(\alpha^2 - \beta^2)$, where $\alpha, \beta \in k$. We claim that $G = \mathbb{F}_q$, which will complete the proof.

If q is odd, then each $w \in \mathbb{F}_q$ is a difference of squares:

$$w = [\tfrac{1}{2}(w+1)]^2 - [\tfrac{1}{2}(w-1)]^2.$$

Hence, if $u \in \mathbb{F}_q$, there are $\alpha, \beta \in \mathbb{F}_q$ with $r^{-1}u = \alpha^2 - \beta^2$, and so $u = r(\alpha^2 - \beta^2) \in G$; therefore, $G = \mathbb{F}_q$. If $q = 2^m$, then the function $u \mapsto u^2$ is an injection $\mathbb{F}_q \to \mathbb{F}_q$ (for if $u^2 = v^2$, then $0 = u^2 - v^2 = (u-v)^2$, and $u = v$). It follows from Exercise 1.58 on page 36 (an injection from a finite set to itself must be a bijection) that this function is surjective, and so every element u has a square root in $\mathbb{F}_q$. In particular, there is $\alpha \in \mathbb{F}_q$ with $r^{-1}u = \alpha^2$, and $u = r\alpha^2 \in G$. •

We need a short technical lemma before giving the main result.

Lemma 5.65. *Let H be a normal subgroup of* $\mathrm{SL}(2, \mathbb{F}_q)$. *If $A \in H$ is similar to*

$$R = \begin{bmatrix} \alpha & \beta \\ \gamma & \delta \end{bmatrix},$$

where $R \in \mathrm{GL}(2, \mathbb{F}_q)$, then there is $u \in \mathbb{F}_q$ so that H contains

$$\begin{bmatrix} \alpha & u^{-1}\beta \\ u\gamma & \delta \end{bmatrix}.$$

Proof. By hypothesis, there is a matrix $P \in \mathrm{GL}(2, \mathbb{F}_q)$ with $R = PAP^{-1}$. There is a matrix $U \in \mathrm{SL}$ and a diagonal matrix $D = \mathrm{diag}\{1, u\}$ with $P^{-1} = UD$, by Lemma 5.60. Therefore, $A = UDRD^{-1}U^{-1}$; since $H \lhd \mathrm{SL}$, we have $DRD^{-1} = U^{-1}AU \in H$. But

$$DRD^{-1} = \begin{bmatrix} 1 & 0 \\ 0 & u \end{bmatrix} \begin{bmatrix} \alpha & \beta \\ \gamma & \delta \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & u^{-1} \end{bmatrix} = \begin{bmatrix} \alpha & u^{-1}\beta \\ u\gamma & \delta \end{bmatrix}. \quad \bullet$$

The next theorem was proved by C. Jordan in 1870 for q prime. In 1893, after F. Cole had discovered a simple group of order 504, E. H. Moore recognized Cole's group as $\mathrm{PSL}(2, \mathbb{F}_8)$, and he then proved the simplicity of $\mathrm{PSL}(2, \mathbb{F}_q)$ for all prime powers $q \geq 4$. We can define $\mathrm{PSL}(m, \mathbb{F}_q)$ for all $m \geq 3$ as $\mathrm{SL}(m, \mathbb{F}_q)/\mathrm{SZ}(m, \mathbb{F}_q)$, and Jordan proved, for all $m \geq 3$, that $\mathrm{PSL}(m, \mathbb{F}_p)$ is simple for all primes p. In 1897, L. E. Dickson proved that $\mathrm{PSL}(m, \mathbb{F}_q)$ is simple for all prime powers q.

We are going to use Corollary 3.101: Two $n \times n$ matrices A and B over a field k are similar (that is, there exists a nonsingular matrix P with $B = PAP^{-1}$) if and only if they both arise from a single linear transformation $\varphi \colon k^n \to k^n$ relative to two choices of bases of k^n. Of course, two nonsingular $n \times n$ matrices A and B over a field k are similar if and only if they are conjugate elements in the group $\mathrm{GL}(n, k)$.

Theorem 5.66 (Jordan–Moore). *The groups* $\mathrm{PSL}(2, \mathbb{F}_q)$ *are simple for all prime powers* $q \geq 4$.

Remark. By Proposition 5.63, $|\mathrm{PSL}(2, \mathbb{F}_2)| = 6$ and $|\mathrm{PSL}(2, \mathbb{F}_3)| = 12$, so that neither of these groups is simple.

It is true that $\mathrm{PSL}(2, k)$ is a simple group for every infinite field k. ◂

Proof. It suffices to prove that a normal subgroup H of $\mathrm{SL}(2, \mathbb{F}_q)$ that contains a matrix not in the center $\mathrm{SZ}(2, \mathbb{F}_q)$ must be all of $\mathrm{SL}(2, \mathbb{F}_q)$.

Suppose, first, that H contains a matrix

$$A = \begin{bmatrix} \alpha & 0 \\ \beta & \alpha^{-1} \end{bmatrix},$$

where $\alpha \neq \pm 1$; that is, $\alpha^2 \neq 1$. If $B = B_{21}(1)$, then H contains the commutator $BAB^{-1}A^{-1} = B_{21}(1 - \alpha^{-2})$, which is a transvection because $1 - \alpha^{-2} \neq 0$. Therefore, $H = \mathrm{SL}(2, \mathbb{F}_q)$, by Lemma 5.64.

To complete the proof, we need only show that H contains a matrix whose top row is $[\alpha\ 0]$, where $\alpha \neq \pm 1$. By hypothesis, there is some matrix $M \in H$ that is not a scalar matrix. Let $\varphi \colon k^2 \to k^2$ be the linear transformation given by $\varphi(v) = Mv$, where v is a 2×1 column vector. If $\varphi(v) = c_v v$ for all v, where $c_v \in k$, then the matrix $[\varphi]$ relative to any basis of k^2 is a diagonal matrix. In this case, M is similar to a diagonal matrix $D = \mathrm{diag}\{\alpha, \beta\}$, and Lemma 5.65 says that $D \in H$. Since $M \notin \mathrm{SZ}(2, \mathbb{F}_q)$, we must have $\alpha \neq \beta$. But $\alpha\beta = \det(M) = 1$, and so $\alpha \neq \pm 1$. Therefore, D is a matrix in H of the desired form.

In the remaining case, there is a vector v with $\varphi(v)$ not a scalar multiple of v, and we saw in Example 3.96(ii) that M is similar to a matrix of the form

$$\begin{bmatrix} 0 & -1 \\ 1 & b \end{bmatrix}$$

(the matrix has this form because it has determinant 1). Lemma 5.65 now says that there is some $u \in k$ with

$$D = \begin{bmatrix} 0 & -u^{-1} \\ u & b \end{bmatrix} \in H.$$

If $T = \mathrm{diag}\{\alpha, \alpha^{-1}\}$ (where α will be chosen in a moment), then the commutator

$$V = (TDT^{-1})D^{-1} = \begin{bmatrix} \alpha^2 & 0 \\ ub(\alpha^{-2} - 1) & \alpha^{-2} \end{bmatrix} \in H.$$

We are done if $\alpha^2 \neq \pm 1$; that is, if there is some nonzero $\alpha \in k$ with $\alpha^4 \neq 1$. If $q > 5$, then such an element α exists, for the polynomial $x^4 - 1$ has at most four roots in a field. If $q = 4$, then every $\alpha \in \mathbb{F}_4$ is a root of the equation $x^4 - x$, and so $\alpha \neq 1$ implies $\alpha^4 \neq 1$.

Only the case $q = 5$ remains. The entry b in D shows up in the lower left corner $v = ub(\alpha^{-2} - 1)$ of the commutator V. There are two subcases depending on whether $b \neq 0$ or $b = 0$. In the first subcase, choose $\alpha = 2$ so that $\alpha^{-2} = 4 = \alpha^2$ and $v = (4-1)ub = 3ub \neq 0$. Now H contains $V^2 = B_{21}(-2v)$, which is a transvection because $-2v = -6ub = 4ub \neq 0$. Finally, if $b = 0$, then D has the form

$$D = \begin{bmatrix} 0 & -u^{-1} \\ u & 0 \end{bmatrix}.$$

Conjugating D by $B_{12}(y)$ for $y \in \mathbb{F}_5$ gives a matrix $B_{12}(y)DB_{12}(-y) \in H$ whose top row is

$$[uy \quad -uy^2 - u^{-1}].$$

If we choose $y = 2u^{-1}$, then the top row is [2 0], and the proof is complete. •

Here are the first few orders of these simple groups:

$$\begin{aligned} |\operatorname{PSL}(2, \mathbb{F}_4)| &= 60; \\ |\operatorname{PSL}(2, \mathbb{F}_5)| &= 60; \\ |\operatorname{PSL}(2, \mathbb{F}_7)| &= 168; \\ |\operatorname{PSL}(2, \mathbb{F}_8)| &= 504; \\ |\operatorname{PSL}(2, \mathbb{F}_9)| &= 360; \\ |\operatorname{PSL}(2, \mathbb{F}_{11})| &= 660. \end{aligned}$$

It can be shown that there is no nonabelian simple group whose order lies between 60 and 168. Indeed, these are all the nonabelian simple groups of order less than 1000.

Some of the orders in the table, namely, 60 and 360, coincide with orders of alternating groups. There do exist nonisomorphic simple groups of the same order; for example, A_8 and $\operatorname{PSL}(3, \mathbb{F}_4)$ are nonisomorphic simple groups of order $\frac{1}{2}8! = 20{,}160$. The next result shows that any two simple groups of order 60 are isomorphic [Exercise 5.53 on page 296 shows that $\operatorname{PSL}(2, \mathbb{F}_9) \cong A_6$].

Proposition 5.67. *If G is a simple group of order 60, then $G \cong A_5$.*

Proof. It suffices to show that G has a subgroup H of index 5, for then Theorem 2.88, the representation on the cosets of H, provides a homomorphism $\varphi : G \to S_5$ with $\ker\varphi \leq H$. As G is simple, the proper normal subgroup $\ker\varphi$ is equal to $\{1\}$, and so G is isomorphic to a subgroup of S_5 of order 60. By Exercise 2.94(ii) on page 114, A_5 is the only subgroup of S_5 of order 60, and so $G \cong A_5$.

Suppose that P and Q are Sylow 2-subgroups of G with $P \cap Q \neq \{1\}$; choose $x \in P \cap Q$ with $x \neq 1$. Now P has order 4, hence is abelian, and so $4 \mid |C_G(x)|$, by Lagrange's theorem. Indeed, since both P and Q are abelian, the subset $P \cup Q$ is contained in $C_G(x)$, so that $|C_G(x)| \geq |P \cup Q| > 4$. Therefore, $|C_G(x)|$ is a proper multiple of 4 which is also a divisor of 60: either $|C_G(x)| = 12$, $|C_G(x)| = 20$, or $|C_G(x)| = 60$. The second case

cannot occur lest $C_G(x)$ have index 3, and representing G on its cosets would show that G is isomorphic to a subgroup of S_3; the third case cannot occur lest $x \in Z(G) = \{1\}$. Therefore, $C_G(x)$ is a subgroup of G of index 5, and we are done in this case. We may now assume that every pair of Sylow 2-subgroups of G intersect in $\{1\}$.

A Sylow 2-subgroup P of G has $r = [G : N_G(P)]$ conjugates, where $r = 3, 5$, or 15. Now $r \neq 3$ (G has no subgroup of index 3). We show that $r = 15$ is not possible by counting elements. Each Sylow 2-subgroup contains three nonidentity elements. Since any two Sylow 2-subgroups intersect trivially (as we saw above), their union contains $15 \times 3 = 45$ nonidentity elements. Now a Sylow 5-subgroup of G must have 6 conjugates (the number r_5 of them is a divisor of 60 satisfying $r_5 \equiv 1 \bmod 5$). But Sylow 5-subgroups are cyclic of order 5, so that the intersection of any pair of them is $\{1\}$, and so the union of them contains $6 \times 4 = 24$ nonidentity elements. We have exceeded the number of elements in G, and so this case cannot occur. •

Corollary 5.68. $\mathrm{PSL}(2, \mathbb{F}_4) \cong A_5 \cong \mathrm{PSL}(2, \mathbb{F}_5)$.

Proof. All three groups are simple and have order 60. •

There are other infinite families of simple matrix groups (in addition to the cyclic groups of prime order, the alternating groups, and the projective unimodular groups), as well as 26 ***sporadic*** simple groups belonging to no infinite family, the largest of which is the "monster" of order approximately 8.08×10^{53}. We refer the interested reader to the books by E. Artin, by R. Carter, and by J. Dieudonné. In fact, all finite simple groups were classified in the 1980's, and an excellent description of this classification can be found in Conway et al, *ATLAS of Finite Groups*.

EXERCISES

5.51 Give a composition series for $\mathrm{GL}(2, \mathbb{F}_5)$ and list its factor groups.

5.52 (i) Prove that $\mathrm{PSL}(2, \mathbb{F}_2) \cong S_3$.

(ii) Prove that $\mathrm{PSL}(2, \mathbb{F}_3) \cong A_4$.

5.53 Prove that $\mathrm{PSL}(2, \mathbb{F}_9) \cong A_6$.

Hint. Let $A = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$ and $B = \begin{bmatrix} 1 & 1+u \\ 0 & 1 \end{bmatrix}$, where $u \in \mathbb{F}_9$ satisfies $u^2 = -1$. If A and B represent elements a and b in $\mathrm{PSL}(2, \mathbb{F}_9)$, prove that ab has order 5 and $|\langle a, b\rangle| = 60$.

5.54 (i) Prove that $\mathrm{SL}(2, \mathbb{F}_5)$ is not solvable.

(ii) Show that a Sylow 2-subgroup of $\mathrm{SL}(2, \mathbb{F}_5)$ is isomorphic to the quaternions $\mathbf{Q}$.

(iii) Prove that the Sylow p-subgroups of $\mathrm{SL}(2, \mathbb{F}_5)$ are cyclic if p is an odd prime. Conclude, for every prime p, that all the Sylow p-subgroups of $\mathrm{SL}(2, \mathbb{F}_5)$ have a unique subgroup of order p.

5.55 Prove that $\mathrm{GL}(2, \mathbb{F}_7)$ is not solvable.

5.56 (i) Prove that $\mathrm{SL}(2, \mathbb{F}_q)$ is the commutator subgroup of $\mathrm{GL}(2, \mathbb{F}_q)$ for all prime powers $q \geq 4$.

(ii) What is the commutator subgroup of $\mathrm{GL}(2, \mathbb{F}_q)$ when $q = 2$ and when $q = 3$?

5.57 Let π be a primitive element of $\mathbb{F}_8$.

(i) What is the order of $A = \begin{bmatrix} \pi & 0 \\ 1 & \pi \end{bmatrix}$ considered as an element of $\mathrm{GL}(2, \mathbb{F}_8)$?

(ii) What is the order of $A = \begin{bmatrix} \pi & 0 & 0 \\ 1 & \pi & 0 \\ 0 & 1 & \pi \end{bmatrix}$ considered as an element of $\mathrm{GL}(3, \mathbb{F}_8)$?

Hint. Show that if $N = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$, then $N^2 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix}$ and $N^3 = 0$, and use the binomial theorem to show that $A^m = \pi^m I + m\pi^{m-1}N + \binom{m}{2}\pi^{m-2}N^2$.

5.5 PRESENTATIONS

How can we describe a group? By Cayley's theorem, a finite group G is isomorphic to a subgroup of the symmetric group S_n, where $n = |G|$, and so G can always be defined as a subgroup of S_n generated by certain permutations. An example of this kind of construction occurs in the following exercise from Carmichael's group theory book[9]:

> Let G be the subgroup of S_{16} generated by the following permutations:
>
> $$\begin{aligned} (a\ c)(b\ d); &\quad (e\ g)(f\ h); \\ (i\ k)(j\ \ell); &\quad (m\ o)(n\ p) \\ (a\ c)(e\ g)(i\ k); &\quad (a\ b)(c\ d)(m\ o); \\ (e\ f)(g\ h)(m\ n)(o\ p); &\quad (i\ j)(k\ \ell). \end{aligned}$$
>
> Prove that $|G| = 256$, $|G'| = 16$,
>
> $$\alpha = (i\ k)(j\ \ell)(m\ o)(n\ p) \in G',$$
>
> but α is not a commutator.

A second way of describing a group is by replacing S_n by $\mathrm{GL}(n, k)$ for some $n \geq 2$ and some field k [remember that all the $n \times n$ permutation matrices form a subgroup of $\mathrm{GL}(n, k)$ isomorphic to S_n, and so every group of order n can be imbedded in $\mathrm{GL}(n, k)$]. We have already described some groups in terms of matrices; for example, we defined the quaternion group $\mathbf{Q}$ in this way. For relatively small groups, descriptions in terms of permutations or matrices are useful, but when n is large, such descriptions are cumbersome.

We can also describe groups as being generated by elements subject to certain relations. For example, the dihedral group D_{2n} could be described as a group of order $2n$ that can be generated by two elements a and b, such that $a^n = 1 = b^2$ and $bab = a^{-1}$. Consider the following definition.

---

[9]Carmichael posed this exercise in the 1930s, before the era of high-speed computers, and he was able to solve it by hand.

Definition. The group of ***generalized quaternions*** $\mathbf{Q}_n$, where $n \geq 3$, is a group of order 2^n that is generated by two elements a and b such that

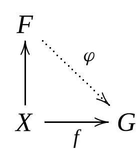
$$a^{2^{n-1}} = 1, \quad bab^{-1} = a^{-1}, \text{ and } \quad b^2 = a^{2^{n-2}}.$$

When $n = 3$, this is the group $\mathbf{Q}$ of order 8. An obvious defect in this definition is that the existence of such a group is left in doubt; for example, is there such a group of order 16? Notice that it is not enough to find a group $G = \langle a, b \rangle$ in which $a^8 = 1$, $bab^{-1} = a^{-1}$, and $b^2 = a^4$. For example, the group $G = \langle a, b \rangle$ in which $a^2 = 1$ and $b = 1$ (which is, of course, cyclic of order 2) satisfies all of the equations.

It was W. von Dyck, in the 1880s, who invented *free groups* in order to make such descriptions rigorous.

Here is a modern definition of a free group.

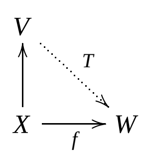
Definition. If X is a subset of a group F, then F is a ***free group*** with ***basis*** X if, for every group G and every function $f : X \to G$, there exists a unique homomorphism $\varphi : F \to G$ with $\varphi(x) = f(x)$ for all $x \in X$.



This definition is modeled on a fundamental result in linear algebra, Theorem 3.92, which is the reason why it is possible to describe linear transformations by matrices.

Theorem. *Let $X = v_1, \ldots, v_n$ be a basis of a vector space V. If W is a vector space and $u_1, \ldots, u_n$ is a list in W, then there exists a unique linear transformation $T : V \to W$ with $T(v_i) = u_i$ for all i.*

We may draw a diagram of this theorem after we note that giving a list $u_1, \ldots, u_n$ of vectors in W is the same thing as giving a function $f : X \to W$, where $f(v_i) = u_i$; after all, a function $f : X \to W$ is determined by its values on $v_i \in X$.



If we knew that free groups exist, then we could define $\mathbf{Q}_n$ as follows. Let F be the free group with basis $X = \{x, y\}$, let R be the normal subgroup of F generated by $\left\{x^{2^{n-1}}, yxy^{-1}x, y^{-2}x^{2^{n-2}}\right\}$, and define $\mathbf{Q}_n = F/R$. It is clear that F/R is a group generated by two elements $a = xR$ and $b = yR$ that satisfy the relations in the definition; what is not clear is that F/R has order 2^n, and this needs proof (see Proposition 5.80).

The first question, then, is whether free groups exist. The idea of the construction is simple and natural, but checking the details is a bit fussy. We begin by describing the ingredients of a free group.

Let X be a nonempty set, and let X^{-1} be a disjoint replica of X; that is, X and X^{-1} are disjoint and there is a bijection $X \to X^{-1}$, which we denote by $x \mapsto x^{-1}$. Define the ***alphabet*** on X to be

$$X \cup X^{-1}.$$

If n is a positive integer, we define a ***word*** on X of ***length*** $n \geq 1$ to be a function $w\colon \{1, 2, \ldots, n\} \to X \cup X^{-1}$. In practice, we shall write a word w of length n as follows: if $w(i) = x_i^{e_i}$, then

$$w = x_1^{e_1} \cdots x_n^{e_n},$$

where $x_i \in X$ and $e_i = \pm 1$. The length n of a word w will be denoted by $|w|$. For example, $|xx^{-1}| = 2$. The ***empty word***, denoted by 1, is a new symbol; the length of the empty word is defined to be 0.

The definition of equality of functions reads here as follows. If $u = x_1^{e_1} \cdots x_n^{e_n}$ and $v = y_1^{d_1} \cdots y_m^{d_m}$ are words, where $x_i, y_j \in X$ for all i, j, then $u = v$ if and only if $m = n$, $x_i = y_i$, and $e_i = d_i$ for all i; thus, every word has a unique spelling.

Definition. A ***subword*** of a word $w = x_1^{e_1} \cdots x_n^{e_n}$ is either the empty word or a word of the form $u = x_r^{e_r} \cdots x_s^{e_s}$, where $1 \leq r \leq s \leq n$. The ***inverse*** of a word $w = x_1^{e_1} \cdots x_n^{e_n}$ is $w^{-1} = x_n^{-e_n} \cdots x_1^{-e_1}$.

It follows that $(w^{-1})^{-1} = w$ for every word w.

The most important words are *reduced* words.

Definition. A word w on X is ***reduced*** if $w = 1$ or if w has no subwords of the form xx^{-1} or $x^{-1}x$, where $x \in X$.

Any two words on X can be multiplied.

Definition. If $u = x_1^{e_1} x_2^{e_2} \cdots x_n^{e_n}$ and $v = y_1^{d_1} \cdots y_m^{d_m}$ are words on X, then their ***juxtaposition*** is the word

$$uv = x_1^{e_1} \cdots x_n^{e_n} y_1^{d_1} \cdots y_m^{d_m}.$$

If 1 is the empty word, then $1v = v$ and $u1 = u$.

Let us try to define a free group as the set of all words on X with operation juxtaposition, with the identity being the empty word 1, and with the inverse of $w = x_1^{e_1} \cdots x_n^{e_n}$ being $w^{-1} = x_n^{-e_n} \cdots x_1^{-e_1}$. There is a problem: If $x \in X$, then we want $x^{-1}x = 1$, but this is not true; $x^{-1}x$ has length 2, not length 0. We can try to remedy this by restricting the elements of F to be reduced words on X; but, even if u and v are reduced, their juxtaposition uv may not be reduced. Of course, we can do all the cancellation to convert uv into a reduced word, but now it is tricky to prove associativity. We solve this problem as follows. Since words such as $zx^{-1}xyzx^{-1}$ and $zyzx^{-1}$, for example, must be identified, it is reasonable to

impose an equivalence relation on the set of all the words on X. If we define the elements of F to be the equivalence classes, then associativity can be proved without much difficulty, and it turns out that there is a unique reduced word in each equivalence class. Therefore, we can regard the elements of F as reduced words and the product of two elements as their juxtaposition followed by reduction.

The casual reader may accept the existence of free groups as just described and proceed to Proposition 5.73 on page 304; here are the details for everyone else.

Definition. Let A and B be words on X, possibly empty, and let $w = AB$. An ***elementary operation*** is either an ***insertion***, changing $w = AB$ to $Aaa^{-1}B$ for some $a \in X \cup X^{-1}$, or a ***deletion*** of a subword of w of the form aa^{-1}, changing $w = Aaa^{-1}B$ to AB.

Definition. We write
$$w \to w'$$
to denote w' arising from w by an elementary operation. Two words u and v on X are ***equivalent***, denoted by $u \sim v$, if there are words $u = w_1, w_2, \ldots, w_n = v$ and elementary operations
$$u = w_1 \to w_2 \to \cdots \to w_n = v.$$
Denote the equivalence class of a word w by $[w]$.

Note that $xx^{-1} \sim 1$ and $x^{-1}x \sim 1$; that is, $[xx^{-1}] = [1] = [x^{-1}x]$.
We construct free groups in two stages.

Definition. A ***semigroup*** is a set having an associative operation; a ***monoid*** is a semigroup S having an identity element 1; that is, $1s = s = s1$ for all $s \in S$. If S and S' are semigroups, then a ***homomorphism*** is a function $f : S \to S'$ such that $f(xy) = f(x)f(y)$; if S and S' are monoids, then a ***homomorphism*** $f : S \to S'$ must also satisfy $f(1) = 1$.

Of course, every group is a monoid, and a homomorphism between groups is a homomorphism of them *qua* monoids.

Example 5.69.
(i) The set of natural numbers $\mathbb{N}$ is a commutative monoid under addition.

(ii) A direct product of monoids is again a monoid (with cooordinatewise operation). In particular, the set $\mathbb{N}^n$ of all n-tuples of natural numbers is a commutative additive monoid. ◂

Here is an example of a noncommutative monoid.

Lemma 5.70. *Let X be a set, and let $\mathcal{W}(X)$ be the set of all words on X [if $X = \varnothing$, then $\mathcal{W}(X)$ consists of only the empty word].*

(i) *$\mathcal{W}(X)$ is a monoid under juxtaposition.*

(ii) *If $u \sim u'$ and $v \sim v'$, then $uv \sim u'v'$.*

(iii) *If G is a group and $f : X \to G$ is a function, then there is a homomorphism $\widetilde{f} : \mathcal{W}(X) \to G$ extending f such that $w \sim w'$ implies $\widetilde{f}(w) = \widetilde{f}(w')$ in G.*

Proof. (i) Associativity of juxtaposition is obvious once we note that there is no cancellation in $\mathcal{W}(X)$.

(ii) The elementary operations that take u to u', when applied to the word uv, give a chain taking uv to $u'v$; the elementary operations that take v to v', when applied to the word $u'v$, give a chain taking $u'v$ to $u'v'$. Hence, $uv \sim u'v'$.

(iii) If $w = x_1^{e_1} \cdots x_n^{e_n}$, then define

$$\widetilde{f}(w) = f(x_1)^{e_1} f(x_2)^{e_2} \cdots f(x_n)^{e_n}.$$

That w has a unique spelling shows that $\widetilde{f}$ is a well-defined function, and it is obvious that $\widetilde{f} : \mathcal{W}(X) \to G$ is a homomorphism.

Let $w \sim w'$. We prove, by induction on the number of elementary operations in a chain from w to w', that $\widetilde{f}(w) = \widetilde{f}(w')$ in G. Consider the deletion $w = Aaa^{-1}B \to AB$, where A and B are subwords of w. That $\widetilde{f}$ is a homomorphism gives

$$\widetilde{f}(Aaa^{-1}B) = \widetilde{f}(A)\,\widetilde{f}(a)\,\widetilde{f}(a)^{-1}\,\widetilde{f}(B).$$

But

$$\widetilde{f}(A)\,\widetilde{f}(a)\,\widetilde{f}(a)^{-1}\,\widetilde{f}(B) = \widetilde{f}(A)\,\widetilde{f}(B) \text{ in } G,$$

because there is cancellation in the group G, so that $\widetilde{f}(Aaa^{-1}B) = \widetilde{f}(AB)$. A similar argument holds for insertions. •

The next proposition will be used to prove that each element in a free group has a normal form.

Proposition 5.71. *Every word w on a set X is equivalent to a unique reduced word.*

Proof. If $X = \varnothing$, then there is only one word on X, the empty word 1, and 1 is reduced.

If $X \neq \varnothing$, we show first that there exists a reduced word equivalent to w. If w has no subword of the form aa^{-1}, where $a \in X \cup X^{-1}$, then w is reduced. Otherwise, delete the first such pair, producing a new word w_1, which may be empty, with $|w_1| < |w|$. Now repeat: If w_1 is reduced, stop; if there is a subword of w_1 of the form aa^{-1}, then delete it, producing a shorter word w_2. Since the lengths are strictly decreasing, this process ends with a reduced word that is equivalent to w.

To prove uniqueness, suppose, on the contrary, that u and v are distinct reduced words and there is a chain of elementary operations

$$u = w_1 \to w_2 \to \cdots \to w_n = v;$$

we may assume that n is minimal. Since u and v are both reduced, the first elementary operation is an insertion, while the last elementary operation is a deletion, and so there must

be a first deletion, say, $w_i \to w_{i+1}$. Thus, the elementary operation $w_{i-1} \to w_i$ inserts aa^{-1} while the elementary operation $w_i \to w_{i+1}$ deletes bb^{-1}, where $a, b \in X \cup X^{-1}$.

There are three cases. If the subwords aa^{-1} and bb^{-1} of w_i coincide, then $w_{i-1} = w_{i+1}$, for w_{i+1} is obtained from w_{i-1} by first inserting aa^{-1} and then deleting it; hence, the chain

$$u = w_1 \to w_2 \to \cdots \to w_{i-1} = w_{i+1} \to \cdots \to w_n = v$$

is shorter than the original shortest chain. The second case has aa^{-1} and bb^{-1} overlapping subwords of w_i; this can happen in two ways. One way is

$$w_i = Aaa^{-1}b^{-1}C,$$

where A, C are subwords of w_i and $a^{-1} = b$; hence, $a = b^{-1}$ and

$$w_i = Aaa^{-1}aC.$$

Therefore, $w_{i-1} = AaC$, because we are inserting aa^{-1}, and $w_{i+1} = AaC$, because we are deleting $bb^{-1} = a^{-1}a$. Thus, $w_{i-1} = w_{i+1}$, and removing w_i gives a shorter chain. The second way an overlap can happen is $w_i = Aa^{-1}aa^{-1}C$, where $b^{-1} = a$. As in the first way, this leads to $w_{i-1} = w_{i+1}$.

Finally, suppose that the subwords aa^{-1} and bb^{-1} do not overlap:

$$w_i = A'aa^{-1}A''bb^{-1}C \text{ and } w_{i+1} = A'aa^{-1}A''C.$$

Now bb^{-1} became a subword of w_i by an earlier insertion of either bb^{-1} or $b^{-1}b$ to some word $w_{j-1} = XY$ with $j < i$; that is, $w_{j-1} \to w_j$, where $w_j = Xbb^{-1}Y$ or $w_j = Xb^{-1}bY$. In the first instance, the subchain $w_{j-1} \to \cdots \to w_{i+1}$ looks like

$$XY \to Xbb^{-1}Y \to \cdots \to Abb^{-1}C \to A'aa^{-1}A''bb^{-1}C \to A'aa^{-1}A''C,$$

where $A = A'A''$. But we can shorten this chain by not inserting bb^{-1}:

$$XY \to \cdots \to AC \to A'aa^{-1}A''C.$$

The only ways the deletion of bb^{-1} can occur in the second instance is if, in $w_{j-1} = XY$, we have $X = X'b$ or $Y = b^{-1}Y'$. If $X = X'b$, then $w_{j-1} = X'bY$ and $w_j = X'bb^{-1}bY$ (and it will be the subword bb^{-1} that will be deleted by the elementary operation $w_i \to w_{i+1}$). As with the first possibility, we do not need the insertion. In more detail, the chain

$$X'bY \to X'bb^{-1}bY \to \cdots \to Abb^{-1}C \to A'aa^{-1}A''bb^{-1}C \to A'aa^{-1}A''C,$$

where the processes $X' \to A$ and $bY \to C$ involve insertions only, can be shortened by removing the insertion of $b^{-1}b$:

$$X'bY \to \cdots \to AC \to A'aa^{-1}A''C.$$

The second case, $Y = b^{-1}Y'$, is treated in the same way. Therefore, in all cases, we are able to shorten the shortest chain, and so no such chain can exist. •

Theorem 5.72. *If X is a set, then the set F of all equivalence classes of words on X with operation $[u][v] = [uv]$ is a free group with basis $\{[x] : x \in X\}$.*

Moreover, every element in F has a normal form: For each $[u] \in F$, there is a unique reduced word w with $[u] = [w]$.

Proof. If $X = \varnothing$, then $\mathcal{W}(\varnothing)$ consists only of the empty word 1, and so $F = \{1\}$. The reader may show that this is, indeed, a free group on $\varnothing$.

Assume now that $X \neq \varnothing$. We have already seen, in Lemma 5.70(ii), that juxtaposition is compatible with the equivalence relation, and so the operation on F is well-defined. The operation is associative, because of associativity in $\mathcal{W}(X)$:

$$\begin{aligned}[u]([v][w]) &= [u][vw] \\ &= [u(vw)] \\ &= [(uv)w] \\ &= [uv][w] \\ &= ([u][v])[w].\end{aligned}$$

The identity is the class [1], the inverse of $[w]$ is $[w^{-1}]$, and so F is a group.

If $[w] \in F$, then

$$[w] = [x_1^{e_1} \cdots x_n^{e_n}] = [x_1^{e_1}][x_2^{e_2}] \cdots [x_n^{e_n}],$$

where $e_i = \pm 1$ for all i, so that F is generated by X (if we identify each $x \in X$ with $[x]$). It follows from Proposition 5.71 that for every $[w]$, there is a unique reduced word u with $[w] = [u]$.

To prove that F is free with basis X, suppose that $f : X \to G$ is a function, where G is a group. Define $\varphi : F \to G$ by

$$\varphi : [x_1^{e_1}][x_2^{e_2}] \cdots [x_n^{e_n}] \mapsto f(x_1)^{e_1} f(x_2)^{e_2} \cdots f(x_n)^{e_n},$$

where $x_1^{e_1} \cdots x_n^{e_n}$ is reduced. Uniqueness of the reduced expression of a word shows that φ is a well-defined function (which obviously extends f). Take note of the relation of φ to the homomorphism $\widetilde{f} : \mathcal{W}(X) \to G$ in Lemma 5.70: When w is reduced,

$$\varphi([w]) = \widetilde{f}(w).$$

It remains to prove that φ is a homomorphism (if so, it is the unique homomorphism extending f, because the subset X generates F). Let $[u], [v] \in F$, where u and v are reduced words, and let $uv \sim w$, where w is reduced. Now

$$\varphi([u][v]) = \varphi([w]) = \widetilde{f}(w),$$

because w is reduced, and

$$\varphi([u])\varphi([v]) = \widetilde{f}(u)\widetilde{f}(v),$$

because u and v are reduced. Finally, $\widetilde{f}(u)\widetilde{f}(v) = \widetilde{f}(w)$, by Lemma 5.70(iii), and so $\varphi([u][v]) = \varphi([u])\varphi([v])$. •

Remark. There is a less fussy proof of the existence of the free group F with basis a given set X, due to M. Barr (see Montgomery–Ralston, *Selected Papers in Algebra*). We have not given this proof here because it does not describe the elements of F, and this description is often needed when using free groups. ◂

We have proved, for every set X, that there exists a free group that is free with basis X. Moreover, the elements of a free group F on X may be regarded as reduced words and the operation may be regarded as juxtaposition followed by reduction; brackets are no longer used, and the elements $[w]$ of F are written as w.

The free group F with basis X that we have just constructed is generated by X. Are any two free groups with basis X isomorphic?

Proposition 5.73.

(i) *Let X_1 be a basis of a free group F_1 and let X_2 be a basis of a free group F_2. If there is a bijection $f: X_1 \to X_2$, then there is an isomorphism $\varphi: F_1 \to F_2$ extending f.*

(ii) *If F is a free group with basis X, then F is generated by X.*

Proof. (i) The following diagram, in which the vertical arrows are inclusions, will help the reader follow the proof.

$$\begin{array}{ccc} F_1 & \underset{\varphi_2}{\overset{\varphi_1}{\rightleftarrows}} & F_2 \\ \uparrow & & \uparrow \\ X_1 & \underset{f^{-1}}{\overset{f}{\rightleftarrows}} & X_2. \end{array}$$

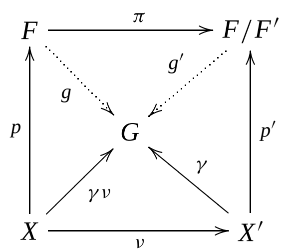
We may regard f as having target F_2, because $X_2 \subseteq F_2$; since F_1 is a free group with basis X_1, there is a homomorphism $\varphi_1: F_1 \to F_2$ extending f. Similarly, there exists a homomorphism $\varphi_2: F_2 \to F_1$ extending f^{-1}. It follows that the composite $\varphi_2\varphi_1: F_1 \to F_1$ is a homomorphism extending 1_X. But the identity 1_{F_1} also extends 1_X, so that uniqueness of the extension gives $\varphi_2\varphi_1 = 1_{F_1}$. In the same way, we see that the other composite $\varphi_1\varphi_2 = 1_{F_2}$, and so φ_1 is an isomorphism.

(ii) Let there be a bijection $f: X_1 \to X$ for some set X_1. If F_1 is the free group with basis X_1 constructed in Theorem 5.72, then X_1 generates F_1. By part (i), there is an isomorphism $\varphi: F_1 \to F$ with $\varphi(X_1) = X$. But if X_1 generates F_1, then $\varphi(X_1)$ generates $\operatorname{im}\varphi$; that is, X generates F. •

There is a notion of rank for free groups, but we must first check that all bases in a free group have the same number of elements.

Lemma 5.74. *If F is a free group with basis $X = x_1, \ldots, x_n$, then F/F' is a free abelian group with basis $X' = x_1F', \ldots, x_nF'$, where F' is the commutator subgroup of F.*

Proof. We begin by noting that X' generates F/F'; this follows from Proposition 5.73(ii), which says that X generates F. We prove that F/F' is a free abelian group with basis X' by using the criterion in Proposition 5.12. Consider the following diagram.



Here, G is an arbitrary abelian group, p and p' are inclusions, π is the natural map, $\nu\colon x \mapsto xF'$, and $\gamma\colon X' \to G$ is a function. Let $g\colon F \to G$ be the unique homomorphism with $gp = \gamma\nu$ given by the definition of free group (for $\gamma\nu\colon X \to G$ is a function), and define $g'\colon F/F' \to G$ by $wF' \mapsto g(w)$ (g' is well-defined because G abelian forces $F' \leq \ker g$). Now $g'p' = \gamma$, for

$$g'p'\nu = g'\pi p = gp = \gamma\nu;$$

since ν is a surjection, it follows that $g'p' = \gamma$. Finally, g' is the unique such map, for if g'' satisfies $g''p' = \gamma$, then g' and g'' agree on the generating set X', hence they are equal. •

Proposition 5.75. *Let F be the free group with basis X. If $|X| = n$, then every basis of F has n elements.*

Proof. By the lemma, F/F' is a free abelian group of rank n. On the other hand, if $y_1, \ldots, y_m$ is another basis of F, then F/F' is a free abelian group of rank m. By Proposition 5.9, we have $m = n$. •

The following definition now makes sense.

Definition. The ***rank*** of a free group F, denoted by $\text{rank}(F)$, is the number of elements in a basis.

Proposition 5.73(i) can now be restated: two free groups of finite rank are isomorphic if and only if they have the same rank.

Proposition 5.76. *Every group G is a quotient of a free group.*

Proof. Let X be a set for which there exists a bijection $f\colon X \to G$ (for example, we could take X to be the underlying set of G and $f = 1_G$), and let F be the free group with basis X. There exists a homomorphism $\varphi\colon F \to G$ extending f, and φ is surjective because f is. Therefore, $G \cong F/\ker\varphi$. •

Let us return to describing groups.

Definition. A ***presentation*** of a group G is an ordered pair

$$G = (X \mid R),$$

where X is a set, R is a set of words on X, and $G = F/N$, where F is the free group with basis X and N is the ***normal subgroup generated by*** R, that is, the subgroup generated by all conjugates of elements of R. We call the set X ***generators***[10] and the set R ***relations***.

Proposition 5.76 says that every group has a presentation.

Definition. A group G is ***finitely generated*** if it has a presentation $(X \mid R)$ with X finite. A group G is called ***finitely presented*** if it has a presentation $(X \mid R)$ in which both X and R are finite.

It is easy to see that a group G is finitely generated if and only if there exists a finite subset $A \subseteq G$ with $G = \langle A \rangle$. There do exist finitely generated groups that are not finitely presented (see my book, *An Introduction to the Theory of Groups*, page 417).

Remark. There are interesting connections between group theory and algebraic topology. If X is a topological space, then its ***fundamental group*** $\pi_1(X)$ is defined to be the set of all homotopy classes of continuous functions $S^1 \to X$, where S^1 is the unit circle. A finite ***simplicial complex*** is a topological space that can be *triangulated* in the sense that it is the union of finitely many vertices, edges, triangles, tetrahedra, and so forth. We can prove that a group G is finitely presented if and only if there is a finite simplicial complex X with $G \cong \pi_1(X)$.

Quite often, a group is known only by some presentation of it. For example, suppose that X is a simplicial complex containing subcomplexes Y_1 and Y_2 such that $Y_1 \cup Y_2 = X$ and $Y_1 \cap Y_2$ is connected. Then ***van Kampen's theorem*** says that a presentation of $\pi_1(X)$ can be given if we know presentations of $\pi_1(Y_1)$ and $\pi_1(Y_2)$. ◂

Example 5.77.
(i) A group has many presentations. For example, $G = \mathbb{I}_6$ has presentations

$$(x \mid x^6)$$

as well as

$$(a, b \mid a^3, b^2, aba^{-1}b^{-1}).$$

A fundamental problem is how to determine whether two presentations give isomorphic groups. It can be proved that no algorithm can exist that solves this problem (see Rotman, *An Introduction to the Theory of Groups*, page 469).

(ii) The free group with basis X has a presentation

$$(X \mid \varnothing).$$

A free group is so called precisely because it has a presentation with no relations. ◂

---

[10] The term *generators* is now being used in a generalized sense, for X is not a subset of G. The subset $\{xN : x \in X\}$ of $G = F/N$ does generate G in the usual sense.

A word on notation. Often, we write the relations in a presentation as equations. Thus, the relations

$$a^3, \quad b^2, \quad aba^{-1}b^{-1}$$

in the second presentation of $\mathbb{I}_6$ may also be written

$$a^3 = 1, \quad b^2 = 1, \quad ab = ba.$$

If r is a word on $x_1, \ldots, x_n$, we may write $r = r(x_1, \ldots, x_n)$. If H is a group and $h_1, \ldots, h_n \in H$, then $r(h_1, \ldots, h_n)$ denotes the element in H obtained from r by replacing each x_i by h_i.

The next, elementary, result is quite useful; we state only the finitely generated case of it.

Theorem 5.78 (von Dyck's Theorem). *Let a group G have a presentation*

$$G = (x_1, \ldots, x_n \mid r_j, \ j \in J);$$

that is, $G = F/N$, where F is free on $\{x_1, \ldots, x_n\}$ and N is the normal subgroup of F generated by all $r_j = r_j(x_1, \ldots, x_n)$. If $H = \langle h_1, \ldots, h_n \rangle$ is a group and if $r_j(h_1, \ldots, h_n) = 1$ in H for all $j \in J$, then there is a surjective homomorphism $G \to H$ with $x_i N \mapsto h_i$ for all i.

Proof. If F is the free group with basis $\{x_1, \ldots, x_n\}$, then there is a homomorphism $\varphi \colon F \to H$ with $\varphi(x_i) = h_i$ for all i. Since $r_j(h_1, \ldots, h_n) = 1$ in H for all $j \in J$, we have $r_j \in \ker\varphi$ for all $j \in J$, which implies $N \leq \ker\varphi$. Therefore, φ induces a (well-defined) homomorphism $G = F/N \to H$ with $x_i N \mapsto h_i$ for all i. •

The next proposition will show how von Dyck's theorem enters into the analysis of presentations, but we begin with the construction of a concrete group of matrices.

Example 5.79.

We are going to construct a group H_n that is a good candidate to be the generalized quaternion group $\mathbf{Q}_n$ for $n \geq 3$ defined on page 298. Consider the complex matrices

$$A = \begin{bmatrix} \omega & 0 \\ 0 & \omega^{-1} \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix},$$

where ω is a primitive 2^{n-1}th root of unity, and let $H_n = \langle A, B \rangle \leq \mathrm{GL}(2, \mathbb{C})$. We claim that A and B satisfy the relations in the definition of the generalized quaternion group. For all $i \geq 1$,

$$A^{2^i} = \begin{bmatrix} \omega^{2^i} & 0 \\ 0 & \omega^{-2^i} \end{bmatrix},$$

so that $A^{2^{n-1}} = I$; indeed, A has order 2^{n-1}. Moreover,

$$B^2 = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix} = A^{2^{n-2}} \quad \text{and} \quad BAB^{-1} = \begin{bmatrix} \omega^{-1} & 0 \\ 0 & \omega \end{bmatrix} = A^{-1}.$$

Notice that A and B do not commute; hence, $B \notin \langle A \rangle$, and so the cosets $\langle A \rangle$ and $B\langle A \rangle$ are distinct. Since A has order 2^{n-1}, it follows that

$$|H_n| \geq |\langle A \rangle \cup B\langle A \rangle| = 2^{n-1} + 2^{n-1} = 2^n.$$

The next theorem will show that $|H_n| = 2^n$. ◂

Proposition 5.80. *For every $n \geq 3$, the generalized quaternion group $\mathbf{Q}_n$ exists.*

Proof. Let G_n be the group defined by the presentation

$$G_n = \big(a, b \mid a^{2^{n-1}} = 1, bab^{-1} = a^{-1}, b^2 = a^{2^{n-2}}\big).$$

The group G_n satisfies all the requirements in the definition of the generalized quaternions with one possible exception: We do not yet know that its order is 2^n. By von Dyck's theorem, there is a surjective homomorphism $G_n \to H_n$, where H_n is the group just constructed in Example 5.79. Hence, $|G_n| \geq 2^n$.

On the other hand, the cyclic subgroup $\langle a \rangle$ in G_n has order at most 2^{n-1}, because $a^{2^{n-1}} = 1$. The relation $bab^{-1} = a^{-1}$ implies that $\langle a \rangle \lhd G_n = \langle a, b \rangle$, so that $G_n/\langle a \rangle$ is generated by the image of b. Finally, the relation $b^2 = a^{2^{n-2}}$ shows that $|G_n/\langle a \rangle| \leq 2$. Hence,

$$|G_n| \leq |\langle a \rangle||G_n/\langle a \rangle| \leq 2^{n-1} \cdot 2 = 2^n.$$

Therefore, $|G_n| = 2^n$, and so $G_n \cong \mathbf{Q}_n$. •

It now follows that the group H_n in Example 5.79 is isomorphic to $\mathbf{Q}_n$.

In Exercise 2.57 on page 81, we gave a concrete construction of the dihedral group D_{2n}, and we can use that group—as in the proof just given— to give a presentation.

Proposition 5.81. *The dihedral group D_{2n} has a presentation*

$$D_{2n} = (a, b \mid a^n = 1, b^2 = 1, bab = a^{-1}).$$

Proof. Let C_{2n} denote the group defined by the presentation, and let D_{2n} be the group of order $2n$ constructed in Exercise 2.57 on page 81. By von Dyck's theorem, there is a surjective homomorphism $f\colon C_{2n} \to D_{2n}$, and so $|C_{2n}| \geq 2n$. To see that f is an isomorphism, we prove the reverse inequality. The cyclic subgroup $\langle a \rangle$ in C_{2n} has order at most n, because $a^n = 1$. The relation $bab^{-1} = a^{-1}$ implies that $\langle a \rangle \lhd C_{2n} = \langle a, b \rangle$, so that $C_{2n}/\langle a \rangle$ is generated by the image of b. Finally, the relation $b^2 = 1$ shows that $|C_{2n}/\langle a \rangle| \leq 2$. Hence,

$$|C_{2n}| \leq |\langle a \rangle||C_{2n}/\langle a \rangle| \leq 2n.$$

Therefore, $|C_{2n}| = 2n$, and so $C_{2n} \cong D_{2n}$. •

In Chapter 2, we classified the groups of order 7 or less. Since groups of prime order are cyclic, it was only a question of classifying the groups of orders 4 and 6. The proof we gave, in Proposition 2.90, that every nonabelian group of order 6 is isomorphic to S_3 was rather complicated, analyzing the representation of a group on the cosets of a cyclic subgroup. Here is a proof in the present spirit.

Proposition 5.82. *If G is a nonabelian group of order* 6*, then $G \cong S_3$.*

Proof. As in the proof of Proposition 2.90, G must contain elements a and b of orders 3 and 2, respectively. Now $\langle a \rangle \lhd G$, because it has index 2, and so either $bab^{-1} = a$ or $bab^{-1} = a^{-1}$. The first possibility cannot occur, because G is not abelian. Therefore, G satisfies the conditions in the presentation of $D_6 \cong S_3$, and so von Dyck's theorem gives a surjective homomorphism $D_6 \to G$. Since both groups have the same order, this map must be an isomorphism. •

We can now classify the groups of order 8.

Theorem 5.83. *Every group G of order* 8 *is isomorphic to*

$$D_8, \quad \mathbf{Q}, \quad \mathbb{I}_8, \quad \mathbb{I}_4 \oplus \mathbb{I}_2, \quad \textit{or} \quad \mathbb{I}_2 \oplus \mathbb{I}_2 \oplus \mathbb{I}_2.$$

Moreover, no two of the displayed groups are isomorphic.

Proof. If G is abelian, then the basis theorem shows that G is a direct sum of cyclic groups, and the fundamental theorem shows that the only such groups are those listed. Therefore, we may assume that G is not abelian.

Now G cannot have an element of order 8, lest it be cyclic, hence abelian; moreover, not every nonidentity element can have order 2, lest G be abelian, by Exercise 2.26 on page 62. We conclude that G must have an element a of order 4; hence, $\langle a \rangle$ has index 2, and so $\langle a \rangle \lhd G$. Choose $b \in G$ with $b \notin \langle a \rangle$; note that $G = \langle a, b \rangle$ because $\langle a \rangle$ has index 2, hence is a maximal subgroup. Now $b^2 \in \langle a \rangle$, because $G/\langle a \rangle$ is a group of order 2, and so $b^2 = a^i$, where $0 \leq i \leq 3$. We cannot have $b^2 = a$ or $b^2 = a^3 = a^{-1}$ lest b have order 8. Therefore, either

$$b^2 = a^2 \quad \text{or} \quad b^2 = 1.$$

Furthermore, $bab^{-1} \in \langle a \rangle$, by normality, and so $bab^{-1} = a$ or $bab^{-1} = a^{-1}$ (for bab^{-1} has the same order as a). Now $bab^{-1} = a$ says that a and b commute, which implies that G is abelian. We conclude that $bab^{-1} = a^{-1}$. Therefore, there are only two possibilities:

$$a^4 = 1, \quad b^2 = a^2, \text{ and } \; bab^{-1} = a^{-1},$$

or

$$a^4 = 1, \quad b^2 = 1, \text{ and } \; bab^{-1} = a^{-1}.$$

By the lemma, the first equations give relations of a presentation for $\mathbf{Q}$, while Proposition 5.81 shows that the second equations give relations of a presentation of D_8. By von Dyck's theorem, there is a surjective homomorphism $\mathbf{Q} \to G$ or $D_8 \to G$; as $|G| = 8$, however, this homomorphism must be an isomorphism.

Finally, Exercise 2.61 on page 82 shows that $\mathbf{Q}$ and D_8 are not isomorphic (for example, $\mathbf{Q}$ has a unique element of order 2 while D_8 has several such elements). •

The reader may continue this classification of the groups G of small order, say, $|G| \leq 15$. Here are the results. By Corollary 2.104, every group of order p^2, where p is a prime, is abelian, and so every group of order 9 is abelian; by the fundamental theorem of finite

abelian groups, there are only two such groups: $\mathbb{I}_9$ and $\mathbb{I}_3 \times \mathbb{I}_3$. If p is a prime, then every group of order $2p$ is either cyclic or dihedral (see Exercise 5.63). Thus, there are only two groups of order 10 and only two groups of order 14. There are 5 groups of order 12, two of which are abelian. The nonabelian groups of order 12 are $D_{12} \cong S_3 \times \mathbb{I}_2$, A_4, and a group T having the presentation

$$T = \big(a, b \mid a^6 = 1, b^2 = a^3 = (ab)^2\big);$$

see Exercise 5.64, which realizes T as a group of matrices. The group[11] T is an example of a *semidirect product*, a construction that will be discussed in Chapter 10. A group of order pq, where $p < q$ are primes and $q \not\equiv 1 \bmod p$, must be cyclic, and so there is only one group of order 15 [see Exercise 10.11(ii) on page 794], There are 14 nonisomorphic groups of order 16, and so this is a good place to stop.

EXERCISES

5.58 Let F be a free group with basis X and let $A \subseteq X$. Prove that if N is the normal subgroup of F generated by A, then F/N is a free group.

5.59 Let F be a free group.

(i) Prove that F has no elements (other than 1) of finite order.

(ii) Prove that a free group F is abelian if and only if $\operatorname{rank}(F) \le 1$.
Hint. Map a free group of rank ≥ 2 onto a nonabelian group.

(iii) Prove that if $\operatorname{rank}(F) \ge 2$, then $Z(F) = \{1\}$, where $Z(F)$ is the center of F,

5.60 Prove that a free group is solvable if and only if it is infinite cyclic (see page 286).

5.61 (i) If G is a finitely generated group and n is a positive integer, prove that G has only finitely many subgroups of index n.
Hint. Consider homomorphisms $G \to S_n$.

(ii) If H and K are subgroups of finite index in a group G, prove that $H \cap K$ also has finite index in G.

5.62 (i) Prove that each of the generalized quaternion groups $\mathbf{Q}_n$ has a unique subgroup of order 2, namely, $\langle b^2 \rangle$, and this subgroup is the center $Z(\mathbf{Q}_n)$.

(ii) Prove that $\mathbf{Q}_n/Z(\mathbf{Q}_n) \cong D_{2^{n-1}}$.

5.63 If p is a prime, prove that every group G of order $2p$ is either cyclic or isomorphic to D_{2p}.
Hint. By Cauchy's theorem, G must contain an element a of order p, and $\langle a \rangle \lhd G$ because it has index 2.

5.64 Let G be the subgroup of $\mathrm{GL}(2, \mathbb{C})$ generated by

$$\begin{bmatrix} \omega & 0 \\ 0 & \omega^2 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 0 & i \\ i & 0 \end{bmatrix},$$

where $\omega = e^{2\pi i/3}$ is a primitive cube root of unity.

(i) Prove that G is a group of order 12 that is not isomorphic to A_4 or to D_{12}.

[11]The group T is called a ***dicyclic group*** of type $(2, 2, 3)$ in Coxeter and Moser, *Generators and Relations for Discrete Groups*, but this terminology in not generally accepted.

(ii) Prove that G is isomorphic to the group T on page 310.

5.65 Prove that every finite group is finitely presented.

5.66 Compute the order of the group G with the presentation

$$G = \big(a, b, c, d \mid bab^{-1} = a^2, bdb^{-1} = d^2, c^{-1}ac = b^2, dcd^{-1} = c^2, bd = db\big).$$

5.67 If X is a nonempty set, define $\Omega(X)$ to be the set of all *positive* words w on X; that is, $\Omega(X)$ is the subset of $\mathcal{W}(X)$ consisting of all $x_1^{e_1} \cdots x_n^{e_n}$ with all $e_i = 1$. Define a ***free monoid***, and prove that $\Omega(X)$ is the free monoid with basis X.

5.6 THE NIELSEN–SCHREIER THEOREM

We are now going to prove one of the most fundamental results about free groups: Every subgroup is also free. This theorem was first proved by J. Nielsen, in 1921, for finitely generated subgroups; the finiteness hypothesis was removed by O. Schreier, in 1926, and so the theorem is now called the Nielsen–Schreier theorem. Nielsen's method actually provides an algorithm, analogous to Gaussian elimination in linear algebra, which replaces a generating set A of a free group F with a basis of $\langle A \rangle$.[12] In particular, if S is a finitely generated subgroup of a free group F, then Nielsen's algorithm replaces any generating set of S with a basis of S, thereby proving that S is free. For an exposition of this proof, we refer the reader to the book of Lyndon and Schupp, pages 4–13.

A second type of proof was found by R. Baer and F. Levi in 1933. It uses a connection, analogous to the correspondence between Galois groups and intermediate fields, between *covering spaces* $\tilde{X}$ of a topological space X and subgroups of its fundamental group $\pi_1(X)$. In particular, if X is a *graph* (a space constructed of edges and vertices), then it can be shown that every covering space is also a graph. It turns out that $\pi_1(\tilde{X})$ is isomorphic to a subgroup of $\pi_1(X)$. Conversely, given any subgroup $S \le \pi_1(X)$, there exists a covering space $\tilde{X}_S$ of X for which $\pi_1(\tilde{X}_S)$ is isomorphic to S. Moreover, $\pi_1(X)$ is a free group whenever X is a graph. Once all these facts are established, the proof proceed as follows. Given a free group F, there is a graph X (a "bouquet of circles") with $F \cong \pi_1(X)$; given a subgroup $S \le F$, we know that $S \cong \pi_1(\tilde{X}_S)$. But $\tilde{X}_S$ is also a graph, so that $\pi_1(\tilde{X}_S)$, and hence S, is free. There are versions of this proof that avoid topology; for example, there is an exposition of such a proof in my book, *An Introduction to the Theory of Groups*, pages 377–384. Interesting variations of this idea are due to J.-P. Serre, in his book *Trees*, who characterized free groups by their action on trees (trees arise as certain *universal* covering spaces of connected graphs), and by P. J. Higgins, who used groupoids.

We give A. J. Weir's proof ["The Reidemeister-Schreier and Kuroš Subgroup Theorems," *Mathematika* 3 (1956), 47–55] of the subgroup theorem because it requires less preparation than the others. The idea arises from a proof of the Reidemeister–Schreier

[12]This theoretical algorithm has evolved into the *Schreier–Sims algorithm*, an efficient way to compute the order of a subgroup $H \le S_n$ when a generating set of H is given; it also can determine whether a specific permutation lies in H.

theorem, which gives a presentation of a subgroup of a group G in terms of a given presentation of G.

Definition. Let S be a subgroup of a group G. A ***transversal*** ℓ of S in G is a subset of G consisting of exactly one element $\ell(Sb) \in Sb$ from every coset Sb, and with $\ell(S) = 1$.

Let F be a free group with basis X, and let S be a subgroup of F. Given a transversal ℓ of S in F, then for each $x \in X$, both $\ell(Sb)x$ and $\ell(Sbx)$ lie in the coset Sbx, and so

$$t_{Sb,x} = \ell(Sb)x\ell(Sbx)^{-1}$$

lies in S. We are going to prove that if the transversal ℓ is chosen wisely, then the set of all $t_{Sb,x}$ that are not 1 form a basis of S, so that S is free.

Let ℓ be a transversal of a subgroup S of a free group F, let the elements $t_{Sb,x}$ be as above, and define Y to be the free group on symbols $y_{Sb,x}$ so that $y_{Sb,x} \mapsto t_{Sb,x}$ is a bijection. Define $\varphi : Y \to S$ to be the homomorphism with

$$\varphi\colon y_{Sb,x} \mapsto t_{Sb,x} = \ell(Sb)x\ell(Sbx)^{-1}.$$

We begin by defining ***coset functions*** $F \to Y$, one for each coset Sb, which we denote by $u \mapsto u^{Sb}$. These functions are not homomorphisms, and we define them all simultaneously by induction on $|u| \geq 0$, where u is a reduced word on X. For all $x \in X$ and all cosets Sb, define

$$1^{Sb} = 1, \quad x^{Sb} = y_{Sb,x}, \quad \text{and} \quad (x^{-1})^{Sb} = (x^{Sbx^{-1}})^{-1}.$$

If $u = x^{\varepsilon}v$ is a reduced word of length $n+1$, where $\varepsilon = \pm 1$ and $|v| = n$, define

$$u^{Sb} = (x^{\varepsilon})^{Sb}v^{Sbx^{\varepsilon}}.$$

Lemma 5.84.

(i) *For all $u, v \in F$, the coset functions satisfy $(uv)^{Sb} = u^{Sb}v^{Sbu}$.*

(ii) *For all $u \in F$, $(u^{-1})^{Sb} = (u^{Sbu^{-1}})^{-1}$.*

(iii) *If $\varphi\colon Y \to S$ is the homomorphism $\varphi : y_{Sb,x} \mapsto t_{Sb,x} = \ell(Sb)x\ell(Sbx)^{-1}$, then, for all $u \in F$, $\varphi(u^{Sb}) = \ell(Sb)u\ell(Sbu)^{-1}$.*

(iv) *The function $\theta\colon S \to Y$, given by $\theta\colon u \mapsto u^S$, is a homomorphism, and $\varphi\theta = 1_S$.*

Proof. (i) The proof is by induction on $|u|$, where u is reduced. If $|u| = 0$, then $u = 1$ and $(uv)^{Sb} = v^{Sb}$; on the other hand, $1^{Sb}v^{Sb1} = v^{Sb}$.

For the inductive step, write $u = x^{\varepsilon}w$. Then

$$\begin{aligned}
(uv)^{Sb} &= (x^{\varepsilon})^{Sb}(wv)^{Sbx^{\varepsilon}} && \text{(definition of coset functions)}\\
&= (x^{\varepsilon})^{Sb}w^{Sbx^{\varepsilon}}v^{Sbx^{\varepsilon}w} && \text{(inductive hypothesis)}\\
&= (x^{\varepsilon})^{Sb}w^{Sbx^{\varepsilon}}v^{Sbu}\\
&= (x^{\varepsilon}w)^{Sb}v^{Sbu}\\
&= u^{Sb}v^{Sbu}.
\end{aligned}$$

(ii) The result follows from

$$1 = 1^{Sb} = (u^{-1}u)^{Sb} = (u^{-1})^{Sb}u^{Sbu^{-1}}.$$

(iii) Note that φ does define a homomorphism because Y is the free group with basis all $y_{Sb,x}$. This proof is also an induction on $|u| \geq 0$. First, $\varphi(1^{Sb}) = \varphi(1) = 1$, while $\ell(S)1\ell(S1)^{-1} = 1$.

For the inductive step, write $u = x^{\varepsilon}v$, where u is reduced. Then

$$\begin{aligned}\varphi(u^{Sb}) = \varphi((x^{\varepsilon}v)^{Sb}) &= \varphi((x^{\varepsilon})^{Sb}v^{Sbx^{\varepsilon}}) \\ &= \varphi((x^{\varepsilon})^{Sb})\varphi(v^{Sbx^{\varepsilon}}) \\ &= \varphi((x^{\varepsilon})^{Sb})\ell(Sbx^{\varepsilon})v\ell(Sbx^{\varepsilon}v)^{-1},\end{aligned}$$

the last equation following from the inductive hypothesis. There are now two cases, depending on the sign ε. If $\varepsilon = +1$, then

$$\begin{aligned}\varphi(u^{Sb}) &= \ell(Sb)x\ell(Sbx)^{-1}\ell(Sbx)v\ell(Sbxv)^{-1} \\ &= \ell(Sb)xv\ell(Sbxv)^{-1} \\ &= \ell(Sb)u\ell(Sbu)^{-1}.\end{aligned}$$

If $\varepsilon = -1$, then

$$\begin{aligned}\varphi(u^{Sb}) &= \varphi((y_{Sbx^{-1},x})^{-1})\ell(Sbx^{-1})v\ell(Sbx^{-1}v)^{-1} \\ &= \Big(\ell(Sbx^{-1})x\ell(Sbx^{-1}x)^{-1}\Big)^{-1}\ell(Sbx^{-1})v\ell(Sbx^{-1}v)^{-1} \\ &= \ell(Sb)x^{-1}\ell(Sbx^{-1})^{-1}\ell(Sbx^{-1})v\ell(Sbx^{-1}v)^{-1} \\ &= \ell(Sb)x^{-1}v\ell(Sbx^{-1}v)^{-1} \\ &= \ell(Sb)u\ell(Sbu)^{-1}.\end{aligned}$$

(iv) For $u \in S$, define $\theta\colon S \to Y$ by

$$\theta : u \mapsto u^{S}$$

(of course, θ is the restriction to S of the coset function $u \mapsto u^{Sb}$ when $b = 1$). Now, if $u, v \in S$, then

$$\theta(uv) = (uv)^{S} = u^{S}v^{Su} = u^{S}v^{S} = \theta(u)\theta(v),$$

because $Su = S$ when $u \in S$. Therefore, θ is a homomorphism. Moreover, if $u \in S$, then part (iii) gives

$$\varphi\theta(u) = \varphi(u^{S}) = \ell(S1)u\ell(S1u)^{-1} = u. \quad \bullet$$

Corollary 5.85. *If S is a subgroup of a free group F and if ℓ is a transversal of S in F, then the set of all $t_{Sb,x}$ that are distinct from 1 generates S.*

Proof. Since the composite $\varphi\theta = 1_S$, the function $\varphi\colon Y \to S$ is surjective; hence, the images $t_{Sb,x}$ of the generators $y_{Sb,x}$ of Y generate $\operatorname{im}\varphi = S$. Of course, we may delete any occurrences of 1 from a generating set. •

The next lemma shows that we have a presentation of S, namely,

$$S = (y_{Sb,x}, \text{ all } x \in X, \text{ all cosets } Sb \mid \ell(Sb)^S, \text{ all cosets } Sb).$$

Lemma 5.86. *If ℓ is a transversal of S in F, then $\ker\varphi$ is the normal subgroup of Y generated by all $\ell(Sb)^S$.*

Proof. Let N be the normal subgroup of Y generated by all $\ell(Sb)^S$, and let $K = \ker\varphi$. By Lemma 5.84(iv), $\theta\colon S \to Y$ is a homomorphism with $\varphi\theta = 1_S$ (where $\varphi\colon y_{Sb,x} \mapsto t_{Sb,x}$ and $\theta\colon u \mapsto u^S$). It follows from Exercise 5.72(ii) on page 318 that K is the normal subgroup of Y generated by $\{y^{-1}\rho(y) : y \in Y\}$, where $\rho = \theta\varphi$. By Lemma 5.84(i),

$$\begin{aligned} y_{Sb,x}^{-1}\rho(y_{Sb,x}) &= y_{Sb,x}^{-1}\left(\ell(Sb)x\ell(Sbx)^{-1}\right)^S \\ &= y_{Sb,x}^{-1}\ell(Sb)^S x^{Sb}\left(\ell(Sbx)^{-1}\right)^{Sbx} \\ &= \left(y_{Sb,x}^{-1}\ell(Sb)^S y_{Sb,x}\right)\left(\ell(Sbx)^{-1}\right)^{Sbx}, \end{aligned}$$

for $x^{Sb} = y_{Sb,x}$ is part of the definition of the coset function $u \mapsto u^{Sb}$. Therefore,

$$y_{Sb,x}^{-1}\rho(y_{Sb,x}) = \left(y_{Sb,x}^{-1}\ell(Sb)^S y_{Sb,x}\right)\left(\ell(Sbx)^S\right)^{-1}, \tag{1}$$

because Lemma 5.84(ii) gives $(\ell(Sbx)^{-1})^{Sbx} = (\ell(Sbx)^S)^{-1}$. It follows from Eq. (1) that $y_{Sb,x}^{-1}\rho(y_{Sb,x}) \in N$, and so $K \le N$. For the reverse inclusion, Eq. (1) says that $\ell(Sb)^S \in K$ if and only if $\ell(Sbx)^S \in K$. Therefore, the desired inclusion can be proved by induction on $|\ell(Sb)|$, and so $K = N$, as desired. •

We now choose a special transversal.

Definition. Let F be a free group with basis X and let S be a subgroup of F. A ***Schreier transversal*** is a transversal ℓ with the property that if $\ell(Sb) = x_1^{\varepsilon_1}x_2^{\varepsilon_2}\cdots x_n^{\varepsilon_n}$ is a reduced word, then every initial segment $x_1^{\varepsilon_1}x_2^{\varepsilon_2}\cdots x_k^{\varepsilon_k}$, for $1 \le k \le n$, is also in the transversal.

Lemma 5.87. *A Schreier transversal exists for every subgroup S of F.*

Proof. Define the ***length*** of a coset Sb, denoted by $|Sb|$, to be the minimum length of the elements $sb \in Sb$. We prove, by induction on $|Sb|$, that there is a representative $\ell(Sb) \in Sb$ such that all its initial segments are representatives of cosets of shorter length.

Begin by defining $\ell(S) = 1$. For the inductive step, let $|Sz| = n + 1$ and let $ux^\varepsilon \in Sz$, where $\varepsilon = \pm 1$ and $|ux^\varepsilon| = n + 1$. Now $|Su| = n$, for if its length were $m < n$, it would have a representative v of length m, and then vx^ε would be a representative of Sz of length $< n + 1$. By induction, $b = \ell(Su)$ exists such that every initial segment is also a representative. Define $\ell(Sz) = bx^\varepsilon$. •

Here is the result we have been seeking.

Theorem 5.88 (Nielsen–Schreier). *Every subgroup S of a free group F is free. In fact, if X is a basis of F and if ℓ is a Schreier transversal of S in F, then a basis for S consists of all $t_{Sb,x} = \ell(Sb)x\ell(Sbx)^{-1}$ that are not* 1.

Proof. Recall that $S \cong Y/K$, where Y is the free group with basis all symbols $y_{Sb,x}$ and $K = \ker\varphi$; by Lemma 5.86, K is equal to the normal subgroup generated by all $\ell(Sb)^S$. By Exercise 5.58 on page 310, it suffices to show that K is equal to the normal subgroup T of Y generated by all ***special*** $y_{Sb,x}$; that is, by those $y_{Sb,x}$ for which $\varphi(y_{Sb,x}) = t_{Sb,x} = 1$. Clearly, $T \le K = \ker\varphi$, and so it suffices to prove the reverse inclusion. We prove, by induction on $|\ell(Sv)|$, that $\ell(Sv)^S$ is a word on the special $y_{Sb,x}$. If $|\ell(Sv)| = 0$, then $\ell(Sv) = \ell(S) = 1$, which is a word on the special $y_{Sb,x}$. If $|\ell(Sv)| > 0$, then $\ell(Sv) = ux^\varepsilon$, where $\varepsilon = \pm 1$ and $|u| < |\ell(Sv)|$. Since ℓ is a Schreier transversal, u is also a representative: $u = \ell(Su)$. By Lemma 5.84(i),

$$\ell(Sv)^S = u^S (x^\varepsilon)^{Su}.$$

By induction, u^S is a word on the special $y_{Sb,x}$, and hence $u^S \in T$.

It remains to prove that $(x^\varepsilon)^{Su}$ is a word on the special $y_{Sb,x}$. If $\varepsilon = +1$, then $(x^\varepsilon)^{Su} = x^{Su} = y_{Su,x}$. But $\ell(Sux) = ux$, because $v = ux$ and ℓ is a Schreier transversal, so that

$$\varphi(y_{Su,x}) = t_{Su,x} = \ell(Su)x\ell(Sux)^{-1} = ux(ux)^{-1} = 1.$$

Therefore, $y_{Su,x}$ is special and x^{Su} lies in T. If $\varepsilon = -1$, then the definition of coset functions gives

$$(x^{-1})^{Su} = (x^{Sux^{-1}})^{-1} = (y_{Sux^{-1},x})^{-1}.$$

Hence,

$$\varphi((x^{-1})^{Su}) = (t_{Sux^{-1},x})^{-1} = [\ell(Sux^{-1})x\ell(Sux^{-1}x)]^{-1} = [\ell(Sux^{-1})x\ell(Su)]^{-1}.$$

Since ℓ is a Schreier transversal, we have $\ell(Su) = u$ and $\ell(Sux^{-1}) = \ell(Sv) = v = ux^{-1}$. Hence,

$$\varphi((x^{-1})^{Su}) = [(ux^{-1})xu^{-1}]^{-1} = 1.$$

Therefore, $y_{Sux^{-1},x}$ is special, $(x^{-1})^{Su} \in T$, and the proof is complete. •

Here is a nice application of the Nielsen–Schreier theorem.

Corollary 5.89. *Let F be a free group, and let $u, v \in F$. Then u and v commute if and only if there is $z \in F$ with $u, v \in \langle z \rangle$.*

Proof. Sufficiency is obvious; if both $u, v \in \langle z \rangle$, then they lie in an abelian subgroup, and hence they commute.

Conversely, the Nielsen–Schreier theorem says that the subgroup $\langle u, v \rangle$ is free. On the other hand, the condition that u and v commute says that $\langle u, v \rangle$ is abelian. But an abelian free group is cyclic, by Exercise 5.59(ii) on page 310: therefore, $\langle u, v \rangle \cong \mathbb{Z}$, as desired. •

The next result shows, in contrast to abelian groups, that a subgroup of a finitely generated group need not be finitely generated.

Corollary 5.90. *If F is a free group of* rank 2*, then its commutator subgroup F' is a free group of infinite rank.*

Proof. Let $\{x, y\}$ be a basis of F. Since F/F' is free abelian with basis $\{xF', yF'\}$, by Lemma 5.74, every coset $F'b$ has a unique representative of the form $x^m y^n$, where $m, n \in \mathbb{Z}$; it follows that the transversal choosing $\ell(F'b) = x^m y^n$ is a Schreier transversal, for every subword of $x^m y^n$ is a word of the same form. If $n > 0$, then $\ell(F'y^n) = y^n$, but $\ell(F'y^n x) = xy^n \neq y^n x$. Therefore, there are infinitely many elements $t_{Sy^n,x} = \ell(F'y^n)x\ell(F'y^n x)^{-1} \neq 1$, and so the result follows from the Nielsen-Schreier theorem. •

Even though an arbitrary subgroup of a finitely generated free group need not be finitely generated, a subgroup of finite index must be finitely generated.

Corollary 5.91. *If F is a free group of finite rank n, then every subgroup S of F having finite index j is also finitely generated. In fact,* $\operatorname{rank}(S) = jn - j + 1$.

Proof. Let $X = \{x_1, \ldots, x_n\}$ be a basis of F, and let $\ell = \{\ell(Sb)\}$ be a Schreier transversal. By Theorem 5.88, a basis of S consists of all those elements $t_{Sb,x}$ not equal to 1, where $x \in X$. There are j choices for Sb and n choices for x, and so there are at most jn elements in a basis of S. Therefore, $\operatorname{rank}(S) \le jn$, and so S is finitely generated.

Call an ordered pair (Sb, x) ***trivial*** if $t_{Sb,x} = 1$; that is, if $\ell(Sb)x = \ell(Sbx)$. We will show that there is a bijection ψ between the family of cosets $\{Sb \neq S\}$ and the trivial ordered pairs, so that there are $j - 1$ trivial ordered pairs. It will then follow that

$$\operatorname{rank}(S) = jn - (j - 1) = jn - j + 1.$$

Since $Sb \neq S$, we have $\ell(Sb) = b = ux^{\varepsilon}$; since ℓ is a Schreier transversal, we have $u \in \ell$. Define $\psi(Sb)$ as follows.

$$\psi(Sux^{\varepsilon}) = \begin{cases} (Su, x) & \text{if } \varepsilon = +1; \\ (Sux^{-1}, x) & \text{if } \varepsilon = -1. \end{cases}$$

Note that $\psi(Sux^{\varepsilon})$ is a trivial ordered pair. If $\varepsilon = +1$, then $\ell(Sux) = \ell(Sb) = b = ux$, so that $\ell(Su)x = ux$ and $t_{Su,x} = 1$. If $\varepsilon = -1$, then $\ell(Sbx) = \ell(Sux^{-1}x) = \ell(Su) = u$, so that $\ell(Sb)x = bx = ux^{-1}x = u$ and $t_{Sb,x} = 1$.

To see that ψ is injective, suppose that $\psi(Sb) = \psi(Sc)$, where $b = ux^\varepsilon$ and $c = vy^\eta$; we assume that x, y lie in the given basis of F and that $\varepsilon = \pm 1$ and $\eta = \pm 1$. There are four possibilities, depending on the signs of ε and η.

$$(Su, x) = (Sv, y); (Su, x) = (Svy^{-1}, y); (Sux^{-1}, x) = (Sv, y); (Su, x) = (Svy^{-1}, y).$$

In every case, equality of ordered pairs gives $x = y$. If $(Su, x) = (Sv, x)$, then $Su = Sv$, hence, $Sb = Sux = Svx = Sc$, as desired. If $(Su, x) = (Svx^{-1}, x)$, then $Su = Svx^{-1} = Sc$, and so $\ell(Su) = \ell(Sc) = c$. But $\ell(Su)x = \ell(Sux) = b$, because (Su, x) is a trivial ordered pair. Hence, $b = \ell(Su)x = cx = vx^{-1}x$, contradicting b (as any element of a Schreier transveral) being reduced. A similar contradiction shows that we cannot have $(Sux^{-1}, x) = (Sv, x)$. Finally, if $(Sux^{-1}, x) = S(vx^{-1}, x)$, then $Sb = Sux^{-1} = Svx^{-1} = Sc$.

To see that ψ is surjective, take a trivial ordered pair (Sw, x); that is, $\ell(Sw)x = wx = \ell(Swx)$. Now $w = ux^\varepsilon$, where $u \in \ell$ and $\varepsilon = \pm 1$. If $\varepsilon = +1$, then w does not end with x^{-1}, and $\psi(Swx) = (Sw, x)$. If $\varepsilon = -1$, then w does end with x^{-1}, and so $\psi(Su) = (Sux^{-1}, x) = (Sw, x)$. •

Corollary 5.92. *There exist nonisomorphic finitely generated groups G and H each of which is isomorphic to a subgroup of the other.*

Proof. If G is a free group of rank 2 and H is a free group of rank 3, then $G \ncong H$. Clearly, G is isomorphic to a subgroup of H. On the other hand, the commutator subgroup G' is free of infinite rank, and so G', hence G, contains a free subgroup of rank 3; that is, H is isomorphic to a subgroup of G. •

We are at the very beginning of a rich subject called ***combinatorial group theory***, which investigates how much can be said about a group given a presentation of it. One of the most remarkable results is the unsolvability of the word problem. A group G has a ***solvable word problem*** if it has a presentation $G = (X \mid R)$ for which there exists an algorithm to determine whether an arbitrary word w on X is equal to the identity element in G (if X and R are finite, it can be proved that this property is independent of the choice of presentation). In the late 1950s, P. S. Novikov and W. W. Boone, independently, proved that there exists a finitely presented group G that does not have a solvable word problem (see Rotman, *An Introduction to the Theory of Groups*, Chapter 12). Other problems involve finding presentations for known groups, as we have done for $\mathbf{Q}_n$ and D_{2n}; an excellent reference for such questions is Coxeter–Moser, *Generators and Relations for Discrete Groups*. Another problem is whether a group defined by a presentation is finite or infinite. For example, ***Burnside's problem*** asks whether a finitely generated group G of finite exponent m, that is, $x^m = 1$ for all $x \in G$, must be finite [W. Burnside had proved that if such a group G happens to be a subgroup of $\mathrm{GL}(n, \mathbb{C})$ for some n, then G is finite]. The answer in general, however, is negative; such a group can be infinite. This was first proved, for m odd and large, by P. S. Novikov and S. I. Adyan, in a long and complicated paper. Using a geometric technique involving *van Kampen diagrams* (see Lyndon–Schupp, *Combinatorial Group Theory*, for an introduction to this subject), A. Yu. Ol'shanskii gave a much shorter

and simpler proof. Finally, S. V. Ivanov was able to complete the solution by showing that the presented group can be infinite when m is even and large. Another geometric technique involves a ***Cayley graph*** of a finitely generated group G, which is a graph depending on a given finite generating set; it can be proved that G is free if and only if it has a Cayley graph that is a tree (see Serre, *Trees*). Finally, the interaction between presentations and algorithms is both theoretical and practical. A theorem of G. Higman (see Rotman, *An Introduction to the Theory of Groups*, Chapter 12) states that a finitely generated group G can be imbedded as a subgroup of a finitely presented group H (that is, H has a presentation with a finite number of generators and a finite number of relations) if and only if G is *recursively presented*: there is a presentation of G whose relations can be given by an algorithm. On the practical side, many efficient algorithms solving group-theoretic problems have been implemented; see Sims, *Computation with Finitely Presented Groups*. The first such algorithm was ***coset enumeration*** (see Lyndon–Schupp, *Combinatorial Group Theory*, pages 163–167), which computes the order of a group G, defined by a presentation, provided that $|G|$ is finite (unfortunately, there can be no algorithm to determine, in advance, whether G is finite).

EXERCISES

5.68 Let G be a finitely generated group, and let $H \leq G$ have finite index. Prove that H is finitely generated.

5.69 Prove that if F is free of finite rank $n \geq 2$, then its commutator subgroup F' is free of infinite rank.

5.70 Let G be a finite group that is not cyclic. If $G \cong F/S$, where F is a free group of finite rank, prove that $\text{rank}(S) > \text{rank}(F)$.

5.71 (i) Prove that if G is a finite group generated by two elements a, b having order 2, then $G \cong D_{2n}$ for some $n \geq 2$.

(ii) Let $G = \langle A, B \rangle \leq \text{GL}(2, \mathbb{Q})$, where

$$A = \begin{bmatrix} 0 & 0 \\ 1 & 1 \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} -1 & 1 \\ 0 & 1 \end{bmatrix}.$$

Show that $A^2 = I = B^2$, but that AB has infinite order. (Exercise prefex:modular group gives another example of a group in which the product of two elements of finite order has infinite order.) The group G is usually denoted by D_∞, and it is called the ***infinite dihedral group***.

5.72 Let Y and S be groups, and let $\varphi\colon Y \to S$ and $\theta\colon S \to Y$ be homomorphisms with $\varphi\theta = 1_S$.

(i) If $\rho\colon Y \to Y$ is defined by $\rho = \theta\varphi$, prove that $\rho\rho = \rho$ and $\rho(a) = a$ for every $a \in \operatorname{im}\theta$. (The homomorphism ρ is called a ***retraction***.)

(ii) If K is the normal subgroup of Y generated by all $y^{-1}\rho(y)$ for $y \in Y$, prove that $K = \ker\varphi$.

Hint. Note that $\ker\varphi = \ker\rho$ because θ is an injection. Use the equation $y = \rho(y)(\rho(y)^{-1})y$ for all $y \in Y$.

6

Commutative Rings II

Our main interest in this chapter is the study of polynomials in several variables. As usual, it is simpler to begin by looking at the more general setting—in this case, commutative rings—before getting involved with polynomial rings. It turns out that the nature of the ideals in a commutative ring is important: for example, we have already seen that gcds exist in PIDs and that they are linear combinations, while these properties may not be enjoyed by other commutative rings. Three special types of ideals—prime ideals, maximal ideals, and finitely generated ideals—are the most interesting. A commutative ring is called *noetherian* if every ideal is finitely generated, and Hilbert's basis theorem shows that $k[x_1, \ldots, x_n]$, where k is a field, is noetherian. Next, we collect several interesting applications of Zorn's lemma (which is discussed in the Appendix), such as the existence of maximal ideals, a theorem of I. S. Cohen saying that a commutative ring is noetherian if and only if every prime ideal is finitely generated, the existence and uniqueness of the algebraic closures of fields, the existence of transcendence bases (as well as Lüroth's theorem), and the existence of maximal separable extensions. The next step introduces a geometric viewpoint in which ideals correspond to certain affine subsets called *varieties*; this discussion involves the Nullstellensatz as well as primary decompositions. Finally, the last section introduces the idea of *Gröbner bases*, which extends the division algorithm from $k[x]$ to $k[x_1, \ldots, x_n]$ and which yields a practical algorithm for deciding many problems that can be encoded in terms of polynomials in several variables.

6.1 Prime Ideals and Maximal Ideals

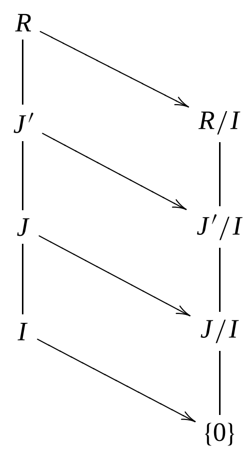
A great deal of the number theory we have presented involves divisibility: Given two integers a and b, when does $a \mid b$; that is, when is a a divisor of b? This question translates into a question about principal ideals, for $a \mid b$ if and only if $(b) \subseteq (a)$. We now introduce two especially interesting types of ideals: *prime ideals*, which are related to Euclid's lemma, and *maximal ideals*.

Let us begin with the analog of Theorem 2.76, the correspondence theorem for groups.

Proposition 6.1 (Correspondence Theorem for Rings). *If I is a proper ideal in a commutative ring R, then there is an inclusion-preserving bijection φ from the set of all intermediate ideals J containing I, that is, $I \subseteq J \subseteq R$, to the set of all the ideals in R/I, given by*

$$\varphi\colon J \mapsto \pi(J) = J/I = \{a + I : a \in J\},$$

where $\pi : R \to R/I$ is the natural map.



Proof. If we forget its multiplication, the commutative ring R is merely an additive abelian group and its ideal I is a (normal) subgroup. The correspondence theorem for groups, Theorem 2.76, now applies, and it gives an inclusion-preserving bijection

$$\Phi : \{\text{all subgroups of } R \text{ containing } I\} \to \{\text{all subgroups of } R/I\},$$

where $\Phi(J) = \pi(J) = J/I$.

If J is an ideal, then $\Phi(J)$ is also an ideal, for if $r \in R$ and $a \in J$, then $ra \in J$, and so

$$(r + I)(a + I) = ra + I \in J/I.$$

Let φ be the restriction of Φ to the set of intermediate ideals; φ is an injection because Φ is a bijection. To see that φ is surjective, let J^* be an ideal in R/I. Now $\pi^{-1}(J^*)$ is an intermediate ideal in R [it contains $I = \pi^{-1}(\{0\})$], and $\varphi(\pi^{-1}(J^*)) = \pi(\pi^{-1}(J^*)) = J^*$, by Proposition 1.50(ii). •

In practice, the correspondence theorem is invoked, tacitly, by saying that every ideal in the quotient ring R/I has the form J/I for some unique ideal J with $I \subseteq J \subseteq R$.

Example 6.2.
Let $I = (m)$ be a nonzero ideal in $\mathbb{Z}$. If J is an ideal in $\mathbb{Z}$ containing I, then $J = (a)$ for some $a \in \mathbb{Z}$, because $\mathbb{Z}$ is a PID, and $(m) \subseteq (a)$ if and only if $a \mid m$. The correspondence theorem now shows that every ideal in the ring $\mathbb{I}_m$ has the form $([a])$ for some divisor a of m, for $J/I = ([a])$. ◄

Definition. An ideal I in a commutative ring R is called a ***prime ideal*** if it is a proper ideal, that is, $I \neq R$, and $ab \in I$ implies $a \in I$ or $b \in I$.

Example 6.3.
(i) Recall that a nonzero commutative ring R is a domain if and only if $ab = 0$ in R implies $a = 0$ or $b = 0$. Thus, the ideal $(0) = \{0\}$ in R is a prime ideal if and only if R is a domain.

(ii) We claim that the prime ideals in $\mathbb{Z}$ are precisely the ideals (p), where either $p = 0$ or p is a prime. Since m and $-m$ generate the same principal ideal, we may restrict our attention to nonnegative generators. If $p = 0$, then the result follows from item (i), for $\mathbb{Z}$ is a domain. If $p > 0$, we show first that (p) is a proper ideal; otherwise, $1 \in (p)$, and there would be an integer a with $ap = 1$, a contradiction. Next, if $ab \in (p)$, then $p \mid ab$. By Euclid's lemma, either $p \mid a$ or $p \mid b$; that is, either $a \in (p)$ or $b \in (p)$. Therefore, (p) is a prime ideal.

Conversely, if $m > 1$ is not a prime, then it has a factorization $m = ab$ with $0 < a < m$ and $0 < b < m$; thus, neither a nor b is a multiple of m, and so neither lies in (m). But $ab = m \in (m)$, and so (m) is not a prime ideal. ◀

Proposition 6.4. *An ideal I in a commutative ring R is a prime ideal if and only if R/I is a domain.*

Proof. Let I be a prime ideal. Since I is a proper ideal, we have $1 \notin I$ and so $1+I \neq 0+I$ in R/I. If $0 = (a+I)(b+I) = ab+I$, then $ab \in I$. Since I is a prime ideal, either $a \in I$ or $b \in I$; that is, either $a + I = 0$ or $b + I = 0$. Hence, R/I is a domain. The converse is just as easy. •

The characterization of prime numbers in Example 6.3(ii) extends to polynomials with coefficients in a field.

Proposition 6.5. *If k is a field, then a nonzero polynomial $p(x) \in k[x]$ is irreducible if and only if $(p(x))$ is a prime ideal.*

Proof. Suppose that $p(x)$ is irreducible. First, (p) is a proper ideal; otherwise, $k[x] = (p)$ and hence $1 \in (p)$, so there is a polynomial $f(x)$ with $1 = p(x)f(x)$. But $p(x)$ has degree at least 1, whereas

$$0 = \deg(1) = \deg(pf) = \deg(p) + \deg(f) \geq \deg(p) \geq 1.$$

This contradiction shows that (p) is a proper ideal. Second, if $ab \in (p)$, then $p \mid ab$, and so Euclid's lemma in $k[x]$ gives $p \mid a$ or $p \mid b$. Thus, $a \in (p)$ or $b \in (p)$. It follows that (p) is a prime ideal.

Conversely, if $(p(x))$ is a prime ideal, then $fg \in (p)$ implies $f \in (p)$ or $g \in (p)$; that is, $p \mid f$ or $p \mid g$. Therefore, Euclid's lemma holds for p, and Exercise 3.31 on page 142 shows that p is irreducible. •

If I is an ideal in a commutative ring R, we may write $I \subsetneq R$ if I is a proper ideal. More generally, if I and J are ideals, we may write $I \subsetneq J$ if $I \subseteq J$ and $I \neq J$.

Here is a second interesting type of ideal.

Definition. An ideal I in a commutative ring R is a ***maximal ideal*** if it is a proper ideal and there is no ideal J with $I \subsetneq J \subsetneq R$.

Thus, if I is a maximal ideal in a commutative ring R and if J is a proper ideal with $I \subseteq J$, then $I = J$. Does every commutative ring R contain a maximal ideal? The (positive) answer to this question involves *Zorn's lemma*, which we will discuss in Section 6.4.

Example 6.6.
The ideal $\{0\}$ is a maximal ideal in a commutative ring R if and only if R is a field. It is shown in Example 3.51(ii) that every nonzero ideal I in R is equal to R itself if and only if every nonzero element in R is a unit. That is, $\{0\}$ is a maximal ideal if and only if R is a field. ◂

Proposition 6.7. *A proper ideal I in a nonzero commutative ring R is a maximal ideal if and only if R/I is a field.*

Proof. The correspondence theorem for rings shows that I is a maximal ideal if and only if R/I has no ideals other than $\{0\}$ and R/I itself; Example 6.6 shows that this property holds if and only if R/I is a field. (Note that since $1 \neq 0$ in a field, I must be a proper ideal.) •

Corollary 6.8. *Every maximal ideal I in a commutative ring R is a prime ideal.*

Proof. If I is a maximal ideal, then R/I is a field. Since every field is a domain, R/I is a domain, and so I is a prime ideal. •

The prime ideals in the polynomial ring $k[x_1, \ldots, x_n]$ can be quite complicated, but when k is an algebraically closed field, Theorem 6.101 shows that every maximal ideal has the form $(x_1 - a_1, \ldots, x_n - a_n)$ for some point $(a_1, \ldots, a_n) \in k^n$; that is, when k is algebraically closed, there is a bijection between k^n and the set of all maximal ideals in $k[x_1, \ldots, x_n]$.

Example 6.9.
The converse of Corollary 6.8 is false. For example, consider the principal ideal (x) in $\mathbb{Z}[x]$. By Exercise 3.83 on page 196, we have

$$\mathbb{Z}[x]/(x) \cong \mathbb{Z};$$

since $\mathbb{Z}$ is a domain, (x) is a prime ideal; since $\mathbb{Z}$ is not a field, (x) is not a maximal ideal. It is not difficult to exhibit a proper ideal J strictly containing (x); let

$$J = \{f(x) \in \mathbb{Z}[x] : f(x) \text{ has even constant term}\}.$$

Since $\mathbb{Z}[x]/J \cong \mathbb{F}_2$ is a field, it follows that J is a maximal ideal containing (x). ◂

Example 6.10.
Let k be a field, and let $a = (a_1, \ldots, a_n) \in k^n$. Define the ***evaluation map***

$$e_a : k[x_1, \ldots, x_n] \to k$$

by

$$e_a : f(x_1, \ldots, x_n) \mapsto f(a) = f(a_1, \ldots, a_n).$$

We have seen, in Example 3.46(iv), that e_a is a surjective ring homomorphism, and so $\ker e_a$ is a maximal ideal. Now $(x_1 - a_1, \ldots, x_n - a_n) \subseteq \ker e_a$. In Exercise 6.6(i) on page 325, however, we shall see that $(x_1 - a_1, \ldots, x_n - a_n)$ is a maximal ideal, and so it must be equal to $\ker e_a$. ◀

The converse of Corollary 6.8 is true when R is a PID.

Theorem 6.11. *If R is a principal ideal domain, then every nonzero prime ideal I is a maximal ideal.*

Proof. Assume that there is a proper ideal J with $I \subseteq J$. Since R is a PID, $I = (a)$ and $J = (b)$ for some $a, b \in R$. Now $a \in J$ implies that $a = rb$ for some $r \in R$, and so $rb \in I$; but I is a prime ideal, so that $r \in I$ or $b \in I$. If $r \in I$, then $r = sa$ for some $s \in R$, and so $a = rb = sab$. Since R is a domain, $1 = sb$, and Exercise 3.18 on page 125 gives $J = (b) = R$, contradicting the hypothesis that J is a proper ideal. If $b \in I$, then $J \subseteq I$, and so $J = I$. Therefore, I is a maximal ideal. •

We can now give a second proof of Proposition 3.116.

Corollary 6.12. *If k is a field and $p(x) \in k[x]$ is irreducible, then the quotient ring $k[x]/(p(x))$ is a field.*

Proof. Since $p(x)$ is irreducible, the principal ideal $I = (p(x))$ is a nonzero prime ideal; since $k[x]$ is a PID, I is a maximal ideal, and so $k[x]/I$ is a field. •

Here are some ways that prime ideals can be used.

Proposition 6.13. *Let P be a prime ideal in a commutative ring R. If I and J are ideals with $IJ \subseteq P$, then $I \subseteq P$ or $J \subseteq P$.*

Proof. Suppose, on the contrary, that $I \not\subseteq P$ and $J \not\subseteq P$; thus, there are $a \in I$ and $b \in J$ with $a, b \notin P$. But $ab \in IJ \subseteq P$, contradicting P being prime. •

The next result is taken from Kaplansky, *Commutative Rings*.

Proposition 6.14. *Let B be a subset of a commutative ring R which is closed under addition and multiplication.*

(i) *Let $J_1, \ldots, J_n$ be ideals in R, at least $n - 2$ of which are prime. If $B \subseteq J_1 \cup \cdots \cup J_n$, then B is contained in some J_i.*

(ii) *Let I be an ideal in R with $I \subsetneq B$. If there are prime ideals $P_1, \ldots, P_n$ such that $B - I \subseteq P_1 \cup \cdots \cup P_n$ (where $B - I$ is the set-theoretic complement of I in B), then $B \subseteq P_i$ for some i.*

Proof. (i) The proof is by induction on $n \geq 2$. For the base step $n = 2$, neither of the ideals J_1 or J_2 need be prime. If $B \not\subseteq J_2$, then there is $b_1 \in B$ with $b_1 \notin J_2$; since $B \subseteq J_1 \cup J_2$, we must have $b_1 \in J_1$. Similarly, if $B \not\subseteq J_1$, there is $b_2 \in B$ with $b_2 \notin J_1$ and $b_2 \in J_2$. However, if $y = b_1 + b_2$, then $y \notin J_1$: otherwise, $b_2 = y - b_1 \in J_1$ (because both y and b_1 are in J_1), a contradiction. Similarly, $y \notin J_2$, contradicting $B \subseteq J_1 \cup J_2$.

For the inductive step, assume that $B \subseteq J_1 \cup \cdots \cup J_{n+1}$, where at least $n-1 = (n+1)-2$ of the J_i are prime ideals. Let

$$D_i = J_1 \cup \ldots \cup \widehat{J_i} \cup \cdots \cup J_{n+1}.$$

Since D_i is a union of n ideals at least $(n-1)-1 = n-2$ of which are prime, the inductive hypothesis allows us to assume that $B \not\subseteq D_i$ for all i. Hence, for all i, there exists $b_i \in B$ with $b_i \notin D_i$; since $B \subseteq D_i \cup J_i$, we must have $b_i \in J_i$. Now $n \geq 3$, so that at least one of the J_i is a prime ideal; for notation, assume that J_1 is prime. Consider the element

$$y = b_1 + b_2 b_3 \cdots b_{n+1}.$$

Since all $b_i \in B$ and B is closed under addition and multiplication, $y \in B$. Now $y \notin J_1$; otherwise, $b_2 b_3 \cdots b_{n+1} = y - b_1 \in J_1$. Since J_1 is prime, some $b_i \in J_1$. This is a contradiction, for $b_i \notin D_i \supseteq J_1$. If $i > 1$ and $y \in J_i$, then $b_2 b_3 \cdots b_{n+1} \in J_i$, because J_i is an ideal, and so $b_1 = y - b_2 b_3 \cdots b_{n+1} \in J_i$. This cannot be, for $b_1 \notin D_1 \supseteq J_i$. Therefore, $y \notin J_i$ for any i, contradicting $B \subseteq J_1 \cup \cdots \cup J_{n+1}$.

(ii) The hypothesis gives $B \subseteq I \cup P_1 \cup \cdots \cup P_n$, so that part (i) gives $B \subseteq I$ or $B \subseteq P_i$. Since I is a proper subset of B, the first possibility cannot occur. •

EXERCISES

6.1 (i) Find all the maximal ideals in $\mathbb{Z}$.

(ii) Find all the maximal ideals in $\mathbb{R}[x]$; that is, describe those $g(x) \in \mathbb{R}[x]$ for which (g) is a maximal ideal.

(iii) Find all the maximal ideals in $\mathbb{C}[x]$.

6.2 Let I be an ideal in a commutative ring R. If J^* and L^* are ideals in R/I, prove that there exist ideals J and L in R containing I such that $J/I = J^*$, $L/I = L^*$, and $(J \cap L)/I = J^* \cap L^*$. Conclude that if $J^* \cap L^* = \{0\}$, then $J \cap L = I$.

Hint. Use the correspondence theorem.

6.3 (i) Give an example of a commutative ring containing two prime ideals P and Q for which $P \cap Q$ is not a prime ideal.

(ii) If $P_1 \supseteq P_2 \supseteq \cdots P_n \supseteq P_{n+1} \supseteq \cdots$ is a decreasing sequence of prime ideals in a commutative ring R, prove that $\bigcap_{n \geq 1} P_n$ is a prime ideal.

6.4 Let $f\colon A \to R$ be a ring homomorphism, where A and R are commutative nonzero rings. Give an example of a prime ideal P in A with $f(P)$ not a prime ideal in R.

6.5 Let $f\colon A \to R$ be a ring homomorphism. If Q is a prime ideal in R, prove that $f^{-1}(Q)$ is a prime ideal in A. Conclude that if J/I is a prime ideal in R/I, where $I \subseteq J \subseteq R$, then J is a prime ideal in R.

6.6 (i) Let k be a field, and let $a_1, \ldots, a_n \in k$. Prove that $(x_1 - a_1, \ldots, x_n - a_n)$ is a maximal ideal in $k[x_1, \ldots, x_n]$.

(ii) Prove that if $x_i - b \in (x_1 - a_1, \ldots, x_n - a_n)$ for some i, where $b \in k$, then $b = a_i$.

(iii) Prove that $\mu\colon k^n \to \{\text{maximal ideals in } k[x_1, \ldots, x_n]\}$, given by

$$\mu\colon (a_1, \ldots, a_n) \mapsto (x_1 - a_1, \ldots, x_n - a_n),$$

is an injection, and give an example of a field k for which μ is not a surjection.

6.7 Prove that if P is a prime ideal in a commutative ring R and if $r^n \in P$ for some $r \in R$ and $n \geq 1$, then $r \in P$.

6.8 Prove that the ideal $(x^2 - 2, y^2 + 1, z)$ in $\mathbb{Q}[x, y, z]$ is a proper ideal.

6.9 (i) Call a nonempty subset S of a commutative ring R ***multiplicatively closed*** if $0 \notin S$ and, if $s, s' \in S$, then $ss' \in S$. Prove that an ideal J which is maximal with the property that $J \cap S = \varnothing$ is a prime ideal. (The existence of such an ideal J is proved, using Zorn's lemma, in Exercise 6.48 on page 374.)

(ii) Let S be a multiplicatively closed subset of a commutative ring R, and suppose that there is an ideal I with $I \cap S = \varnothing$. If P is an ideal maximal such that $I \subseteq P$ and $P \cap S = \varnothing$, prove that P is a prime ideal.

6.10 (i) If I and J are ideals in a commutative ring R, define

$$IJ = \Big\{\text{all finite sums } \sum_{\ell} a_\ell b_\ell \colon a_\ell \in I \text{ and } b_\ell \in J\Big\}.$$

Prove that IJ is an ideal in R and that $IJ \subseteq I \cap J$.

(ii) If $I = (2)$ is the ideal of even integers in $\mathbb{Z}$, prove that $I^2 = IJ \subsetneq I \cap J = I$.

(iii) Let P be a prime ideal and let $Q_1, \ldots\ Q_r$ be ideals. Prove that if $Q_1 \cap \cdots \cap Q_r \subseteq P$, then $Q_i \subseteq P$ for some i.

6.11 Let I and J be ideals in a commutative ring R.

(i) Prove that the map $R/(I \cap J) \to R/I \times R/J$, given by $\varphi\colon r \mapsto (r + I, r + J)$, is an injection.

(ii) Call I and J ***coprime*** if $I + J = R$. Prove that if I and J are coprime, then the ring homomorphism $\varphi\colon R/(I \cap J) \to R/I \times R/J$ in part (i) is a surjection.

Hint. If I and J are coprime, there are $a \in I$ and $b \in J$ with $1 = a + b$. If $r, r' \in R$, prove that$(d + I, d + J) = (r + I, r' + J) \in R/I \times R/J$, where $d = r'a + rb$.

(iii) Generalize the ***Chinese remainder theorem*** as follows. Let R be a commutative ring and let $I_1, \ldots, I_n$ be pairwise coprime ideals; that is, I_i and I_j are coprime for all $i \neq j$. Prove that if $a_1, \ldots, a_n \in R$, then there exists $r \in R$ with $r + I_i = a_i + I_i$ for all i.

6.12 If I and J are coprime ideals in a commutative ring R, prove that

$$I \cap J = IJ.$$

6.13 If I is an ideal in a commutative ring R and if S is a subset of R, define the ***colon ideal***[1]

$$(I : S) = \{r \in R : rs \in I \text{ for all } s \in S\}.$$

(i) Prove that $(I : S)$ is an ideal.

(ii) If $J = (S)$ is the ideal generated by S, prove that $(I : S) = (I : J)$.

(iii) Let R be a domain and let $a, b \in R$, where $b \neq 0$. If $I = (ab)$ and $J = (b)$, prove that $(I : J) = (a)$.

6.14 (i) Let I and J be ideals in a commutative ring R. Prove that $I \subseteq (I : J)$ and that $J(I : J) \subseteq I$.

(ii) Prove that if $I = Q_1 \cap \cdots \cap Q_r$, then

$$(I : J) = (Q_1 : J) \cap \cdots \cap (Q_r : J).$$

(iii) If I is an ideal in a commutative ring R, and if $J = J_1 + \cdots + J_n$ is a sum of ideals, prove that

$$(I : J) = (I : J_1) \cap \cdots \cap (I : J_n).$$

6.15 A ***Boolean ring*** is a commutative ring R in which $a^2 = a$ for all $a \in R$. Prove that every prime ideal in a Boolean ring is a maximal ideal. (See Exercise 8.21 on page 533.)
Hint. When is a Boolean ring a domain?

6.16 A commutative ring R is called a ***local ring*** if it has a unique maximal ideal.

(i) If p is a prime, prove that the ring of ***p-adic fractions***,

$$\mathbb{Z}_{(p)} = \{a/b \in \mathbb{Q} : p \nmid b\},$$

is a local ring.

(ii) If R is a local ring with unique maximal ideal $\mathfrak{m}$, prove that $a \in R$ is a unit if and only if $a \notin \mathfrak{m}$.

Hint. You may assume that every nonunit in a commutative ring lies in some maximal ideal (this result is proved using Zorn's lemma).

6.2 UNIQUE FACTORIZATION DOMAINS

We have proved unique factorization theorems in $\mathbb{Z}$ and in $k[x]$, where k is a field. We are now going to prove a common generalization: Every PID has a unique factorization theorem. We will then prove a theorem of Gauss: If R has unique factorization, then so does $R[x]$. A corollary is that there is unique factorization in the ring $k[x_1, \ldots, x_n]$ of all polynomials in several variables over a field k. One immediate consequence is that any two polynomials in several variables have a gcd.

We begin by generalizing some earlier definitions.

[1]This ideal is also called the ***ideal quotient***.

Definition. Elements a and b in a commutative ring R are ***associates*** if there exists a unit $u \in R$ with $b = ua$.

For example, in $\mathbb{Z}$, the units are ± 1, and so the only associates of an integer m are $\pm m$; in $k[x]$, where k is a field, the units are the nonzero constants, and so the only associates of a polynomial $f(x) \in k[x]$ are the polynomials $uf(x)$, where $u \in k$ and $u \neq 0$. The only units in $\mathbb{Z}[x]$ are ± 1 (see Exercise 6.19 on page 339), and so the only associates of a polynomial $f(x) \in \mathbb{Z}[x]$ are $\pm f(x)$.

In any commutative ring R, associates a and b generate the same principal ideal; the converse may be false if R is not a domain.

Proposition 6.15. *Let R be a domain and let $a, b \in R$.*

(i) *$a \mid b$ and $b \mid a$ if and only if a and b are associates.*

(ii) *The principal ideals (a) and (b) are equal if and only if a and b are associates.*

Proof. (i) If $a \mid b$ and $b \mid a$, there are $r, s \in R$ with $b = ra$ and $a = sb$, and so $b = ra = rsb$. If $b = 0$, then $a = 0$ (because $b \mid a$); if $b \neq 0$, then we may cancel it (R is a domain) to obtain $1 = rs$. Hence, r and s are units, and a and b are associates. The converse is obvious.

(ii) If $(a) = (b)$, then $a \in (b)$; hence, $a = rb$ for some $r \in R$, and so $b \mid a$. Similarly, $b \in (a)$ implies $a \mid b$, and so (i) shows that a and b are associates.

Conversely, if $a = ub$, where u is a unit, then $a \in (b)$ and $(a) \subseteq (b)$. Similarly, $b = u^{-1}a$ implies $(b) \subseteq (a)$, and so $(a) = (b)$. •

Recall that an element p in a domain R is *irreducible* if it is neither 0 nor a unit and if its only factors are units or associates of p. For example, the irreducibles in $\mathbb{Z}$ are the numbers $\pm p$, where p is a prime, and the irreducibles in $k[x]$, where k is a field, are the irreducible polynomials $p(x)$; that is, $\deg(p) \geq 1$ and $p(x)$ has no factorization $p(x) = f(x)g(x)$ where $\deg(f) < \deg(p)$ and $\deg(g) < \deg(p)$. This characterization of irreducible polynomials does not persist in rings $R[x]$ when R is not a field. For example, in $\mathbb{Z}[x]$, the polynomial $f(x) = 2x + 2$ cannot be factored into two polynomials, each having degree smaller than $\deg(f) = 1$, yet $f(x)$ is not irreducible, for in the factorization $2x + 2 = 2(x + 1)$, neither 2 nor $x + 1$ is a unit.

Corollary 6.16. *If R is a* PID *and $p \in R$ is irreducible, then (p) is a prime ideal.*

Proof. Let I be an ideal with $(p) \subseteq I$. Since R is a PID, there is $q \in R$ with $I = (q)$. Hence, $p \in (q)$, and so $p = rq$ for some $r \in R$. Irreducibility of p says that q is either an associate of p or a unit. In the first case, $(p) = (q)$, by Proposition 6.15; in the second case, $(q) = R$. It follows that (p) is a maximal ideal, and hence it is a prime ideal, by Corollary 6.8. •

Here is the definition we have been seeking.

Definition. A domain R is a ***unique factorization domain*** (**UFD**) if

(i) every $r \in R$, neither 0 nor a unit, is a product[2] of irreducibles;

(ii) if $up_1 \cdots p_m = vq_1 \cdots q_n$, where u and v are units and all p_i and q_j are irreducible, then $m = n$ and there is a permutation $\sigma \in S_n$ with p_i and $q_{\sigma(i)}$ associates for all i.

When we proved that $\mathbb{Z}$ and $k[x]$, for k a field, have unique factorization into irreducibles, we did not mention associates because, in each case, irreducible elements were always replaced by favorite choices of associates: In $\mathbb{Z}$, *positive* irreducibles (i.e., primes) are chosen; in $k[x]$, *monic* irreducible polynomials are chosen. The reader should see, for example, that the statement: "$\mathbb{Z}$ is a UFD" is just a restatement of the fundamental theorem of arithmetic.

Proposition 6.17. *Let R be a domain in which every $r \in R$, neither* 0 *nor a unit, is a product of irreducibles. Then R is a* UFD *if and only if (p) is a prime ideal in R for every irrreducible element $p \in R$.*[3]

Proof. Assume that R is a UFD. If $a, b \in R$ and $ab \in (p)$, then there is $r \in R$ with

$$ab = rp.$$

Factor each of a, b, and r into irreducibles; by unique factorization, the left side of the equation must involve an associate of p. This associate arose as a factor of a or b, and hence $a \in (p)$ or $b \in (p)$.

The proof of the converse is merely an adaptation of the proof of the fundamental theorem of arithmetic. Assume that

$$up_1 \cdots p_m = vq_1 \cdots q_n,$$

where p_i and q_j are irreducible elements and u, v are units. We prove, by induction on $\max\{m, n\} \geq 1$, that $n = m$ and the q's can be reindexed so that q_i and p_i are associates for all i. If $\max\{m, n\} = 1$, then $up_1 = vq_1$, $up_1 = v$, or $u = vq_1$. The latter two cases cannot occur, for irreducible elements are not units, and so the base step is true. For the inductive step, the given equation shows that $p_1 \mid q_1 \cdots q_n$. By hypothesis, (p_1) is a prime ideal (which is the analog of Euclid's lemma), and so there is some q_j with $p_1 \mid q_j$. But q_j, being irreducible, has no divisors other than units and associates, so that q_j and p_1 are associates: $q_j = up_1$ for some unit u. Canceling p_1 from both sides, we have $p_2 \cdots p_m = uq_1 \cdots \widehat{q_j} \cdots q_n$. By the inductive hypothesis, $m - 1 = n - 1$ (so that $m = n$), and, after possible reindexing, q_i and p_i are associates for all i. •

The proofs we have given that $\mathbb{Z}$ and $k[x]$, where k is a field, are UFDs involve the division algorithm; as a consequence, it is not difficult to generalize them to prove that every euclidean ring is a UFD. We now show that every PID is, in fact, a UFD; the proof uses a new idea: chains of ideals.

[2]To avoid long phrases, we allow a product of irreducibles to have only one factor; that is, an irreducible element is regarded as a product of irreducibles.

[3]An element p for which (p) is a nonzero prime ideal is often called a ***prime element***. Such elements have the property that $p \mid ab$ implies $p \mid a$ or $p \mid b$.

Lemma 6.18.

(i) *If R is a commutative ring and*

$$I_1 \subseteq I_2 \subseteq \cdots \subseteq I_n \subseteq I_{n+1} \subseteq \cdots$$

is an ascending chain of ideals in R, then $J = \bigcup_{n \geq 1} I_n$ is an ideal in R.

(ii) *If R is a* PID*, then it has no infinite strictly ascending chain of ideals*

$$I_1 \subsetneq I_2 \subsetneq \cdots \subsetneq I_n \subsetneq I_{n+1} \subsetneq \cdots.$$

(iii) *Let R be a* PID*. If $r \in R$ is neither 0 nor a unit, then r is a product of irreducibles.*

Proof. (i) We claim that J is an ideal. If $a \in J$, then $a \in I_n$ for some n; if $r \in R$, then $ra \in I_n$, because I_n is an ideal; hence, $ra \in J$. If $a, b \in J$, then there are ideals I_n and I_m with $a \in I_n$ and $b \in I_m$; since the chain is ascending, we may assume that $I_n \subseteq I_m$, and so $a, b \in I_m$. As I_m is an ideal, $a + b \in I_m$ and, hence, $a + b \in J$. Therefore, J is an ideal.

(ii) If, on the contrary, an infinite strictly ascending chain exists, then define $J = \bigcup_{n \geq 1} I_n$. By (i), J is an ideal; since R is a PID, we have $J = (d)$ for some $d \in J$. Now d got into J by being in I_n for some n. Hence

$$J = (d) \subseteq I_n \subsetneq I_{n+1} \subseteq J,$$

and this is a contradiction.

(iii) A divisor r of an element $a \in R$ is called a *proper divisor* of a if r is neither a unit nor an associate of a. If r is a divisor of a, then $(a) \subseteq (r)$; if r is a proper divisor, then $(a) \subsetneq (r)$, for if the inequality is not strict, then $(a) = (r)$, and this forces a and r to be associates, by Proposition 6.15.

Call a nonzero nonunit $a \in R$ *good* if it is a product of irreducibles; otherwise, call a *bad*. We must show that there are no bad elements. If a is bad, it is not irreducible, and so $a = rs$, where both r and s are proper divisors. But the product of good elements is good, and so at least one of the factors, say r, is bad. The first paragraph shows that $(a) \subsetneq (r)$. It follows, by induction, that there exists a sequence $a_1 = a, a_2 = r, a_3, \ldots, a_n, \ldots$ of bad elements with each a_{n+1} a proper divisor of a_n, and this sequence yields a strictly ascending chain

$$(a_1) \subsetneq (a_2) \subsetneq \cdots \subsetneq (a_n) \subsetneq (a_{n+1}) \subsetneq \cdots,$$

contradicting part (i) of this lemma. •

Theorem 6.19. *If R is a* PID*, then R is a* UFD*. In particular, every euclidean ring is a* UFD.

Proof. In view of the last two results, it suffices to prove that (p) is a prime ideal whenever p is irreducible. Since R is a PID, Proposition 6.16 shows that (p) is a prime ideal. •

The notion of gcd can be defined in any commutative ring R. Example 6.21 shows that there exist domains R containing a pair of elements having no gcd. If $a_1, \ldots, a_n \in R$, then a *common divisor* of $a_1, \ldots, a_n$ is an element $c \in R$ with $c \mid a_i$ for all i. A *greatest common divisor* or *gcd* of $a_1, \ldots, a_n$ is a common divisor d with $c \mid d$ for every common divisor c. Even in the familiar examples of $\mathbb{Z}$ and $k[x]$, gcd's are not unique unless an extra condition is imposed. For example, in $k[x]$, where k is a field, we impose the condition that nonzero gcd's are monic polynomials. In a general PID, however, elements may not have favorite associates.

If R is a domain, then it is easy to see that if d and d' are gcd's of elements $a_1, \ldots, a_n$, then $d \mid d'$ and $d' \mid d$. It follows from Proposition 6.15 that d and d' are associates and, hence, that $(d) = (d')$. Thus, gcd's are not unique, but they all generate the same principal ideal.

The idea in Proposition 1.17 carries over to show that gcd's do exist in UFDs.

Proposition 6.20. *If R is a* UFD*, then a* gcd *of any finite set of elements $a_1, \ldots, a_n$ in R exists.*

Proof. It suffices to prove that a gcd of two elements a and b exists; the general result follows by induction on the number of elements.

There are units u and v and distinct irreducibles $p_1, \ldots, p_t$ with

$$a = up_1^{e_1} p_2^{e_2} \cdots p_t^{e_t}$$

and

$$b = vp_1^{f_1} p_2^{f_2} \cdots p_t^{f_t},$$

where $e_i \geq 0$ and $f_i \geq 0$ for all i. It is easy to see that if $c \mid a$, then the factorization of c into irreducibles is $c = wp_1^{g_1} p_2^{g_2} \cdots p_t^{g_t}$, where w is a unit and $0 \leq g_i \leq e_i$ for all i. Thus, c is a common divisor of a and b if and only if $g_i \leq m_i$ for all i, where

$$m_i = \min\{e_i, f_i\}.$$

It is now clear that $p_1^{m_1} p_2^{m_2} \cdots p_t^{m_t}$ is a gcd of a and b. •

It is not difficult to see that if $a_i = u_i p_1^{e_{i1}} p_2^{e_{i2}} \cdots p_t^{e_{it}}$, where $e_{ij} \geq 0$, u_i are units, and $i = 1, \ldots, n$, then

$$d = p_1^{\mu_1} p_2^{\mu_2} \cdots p_t^{\mu_t}$$

is a gcd of $a_1, \ldots, a_n$, where $\mu_j = \min\{e_{1j}, e_{2j}, \ldots, e_{nj}\}$.

We caution the reader that we have not proved that a gcd of elements $a_1, \ldots, a_n$ is a linear combination of them; indeed, this may not be true (see Exercise 6.21 on page 339).

Example 6.21.
Let k be a field and let R be the subring of $k[x]$ consisting of all polynomials $f(x) \in k[x]$ having no linear term; that is, $f(x) = a_0 + a_2x^2 + \cdots + a_nx^n$. In Exercise 3.60 on page 158, we showed that x^5 and x^6 have no gcd in R. It now follows from Proposition 6.20 that R is not a UFD. [Another example of a domain that is not a UFD is given in Exercise 6.31(ii) on page 340.] ◄

Definition. Elements $a_1, \ldots, a_n$ in a UFD R are called ***relatively prime*** if their gcd is a unit; that is, if every common divisor of $a_1, \ldots, a_n$ is a unit.

We are now going to prove that if R is a UFD, then so is $R[x]$. Recall Exercise 3.21 on page 130: If R is a domain, then the units in $R[x]$ are the units in R.

Definition. A polynomial $f(x) = a_n x^n + \cdots + a_1 x + a_0 \in R[x]$, where R is a UFD, is called ***primitive*** if its coefficients are relatively prime; that is, the only common divisors of $a_n, \ldots, a_1, a_0$ are units.

Of course, every monic polynomial is primitive. Observe that if $f(x)$ is not primitive, then there exists an irreducible $q \in R$ that divides each of its coefficients: If the gcd is a nonunit d, then take for q any irreducible factor of d.

Example 6.22.
We now show, for a UFD R, that every irreducible $p(x) \in R[x]$ of positive degree is primitive. If not, then there is an irreducible $q \in R$ with $p(x) = qg(x)$; note that $\deg(q) = 0$ because $q \in R$. Since $p(x)$ is irreducible, its only factors are units and associates, and so q must be an associate of $p(x)$. But every unit in $R[x]$ has degree 0 [i.e., is a constant (for $uv = 1$ implies $\deg(u) + \deg(v) = \deg(1) = 0$)]; hence, associates in $R[x]$ have the same degree. Therefore, q is not an associate of $p(x)$, because the latter has positive degree, and so $p(x)$ is primitive. ◀

We begin with a technical lemma.

Lemma 6.23 (Gauss's Lemma). *If R is a* UFD *and $f(x), g(x) \in R[x]$ are both primitive, then their product $f(x)g(x)$ is also primitive.*

Proof. If $\pi \colon R \to R/(p)$ is the natural map $\pi \colon a \mapsto a + (p)$, then Proposition 3.48 shows that the function $\widetilde{\pi} \colon R[x] \to (R/(p))[x]$, which replaces each coefficient c of a polynomial by $\pi(c)$, is a ring homomorphism. If a polynomial $h(x) \in R[x]$ is not primitive, there is some irreducible p such that all the coefficients of $\widetilde{\pi}(h)$ are 0 in $R/(p)$; that is, $\widetilde{\pi}(h) = 0$ in $(R/(p))[x]$. Thus, if the product $f(x)g(x)$ is not primitive, there is some irreducible p with $0 = \widetilde{\pi}(fg) = \widetilde{\pi}(f)\widetilde{\pi}(g)$ in $(R/(p))[x]$. Since (p) is a prime ideal, $R/(p)$ is a domain, and hence $(R/(p))[x]$ is also a domain. But, neither $\widetilde{\pi}(f)$ nor $\widetilde{\pi}(g)$ is 0 in $(R/(p))[x]$, because f and g are primitive, and this contradicts $(R/(p))[x]$ being a domain. •

Lemma 6.24. *Let R be a* UFD, *let $Q = \mathrm{Frac}(R)$, and let $f(x) \in Q[x]$ be nonzero.*

(i) *There is a factorization*

$$f(x) = c(f) f^*(x),$$

where $c(f) \in Q$ and $f^*(x) \in R[x]$ is primitive. This factorization is unique in the sense that if $f(x) = qg^*(x)$, where $q \in Q$ and $g^*(x) \in R[x]$ is primitive, then there is a unit $w \in R$ with $q = wc(f)$ and $g^*(x) = w^{-1} f^*(x)$.

(ii) *If $f(x), g(x) \in R[x]$, then $c(fg)$ and $c(f)c(g)$ are associates in R and $(fg)^*$ and f^*g^* are associates in $R[x]$.*

(iii) *Let $f(x) \in Q[x]$ have a factorization $f(x) = qg^*(x)$, where $q \in Q$ and $g^*(x) \in R[x]$ is primitive. Then $f(x) \in R[x]$ if and only if $q \in R$.*

(iv) *Let $g^*(x), f(x) \in R[x]$. If $g^*(x)$ is primitive and $g^*(x) \mid bf(x)$, where $b \in R$ and $b \neq 0$, then $g^*(x) \mid f(x)$.*

Proof. (i) Clearing denominators, there is $b \in R$ with $bf(x) \in R[x]$. If d is the gcd of the coefficients of $bf(x)$, then $(b/d)f(x) \in R[x]$ is a primitive polynomial. If we define $c(f) = d/b$ and $f^*(x) = c(f)f(x)$, then $f^*(x)$ is primitive and $f(x) = c(f)f^*(x)$.

To prove uniqueness, suppose that $c(f)f^*(x) = f(x) = qg^*(x)$, where $c(f), q \in Q$ and $f^*(x), g^*(x) \in R[x]$ are primitive. Exercise 6.17 on page 339 allows us to write $q/c(f)$ in lowest terms: $q/c(f) = u/v$, where u and v are relatively prime elements of R. The equation $vf^*(x) = ug^*(x)$ holds in $R[x]$; equating like coefficients, v is a common divisor of each coefficient of $ug^*(x)$. Since u and v are relatively prime, Exercise 6.18(i) on page 339 gives v a common divisor of the coefficients of $g^*(x)$. But $g^*(x)$ is primitive, and so v is a unit. A similar argument shows that u is a unit. Therefore, $q/c(f) = u/v$ is a unit in R, call it w; we have $wc(f) = q$ and $f^*(x) = wg^*(x)$, as desired.

(ii) There are two factorizations of $f(x)g(x)$ in $R[x]$: $f(x)g(x) = c(fg)(f(x)g(x))^*$ and $f(x)g(x) = c(f)f^*(x)c(g)g^*(x) = c(f)c(g)f^*(x)g^*(x)$. Since the product of primitive polynomials is primitive, each of these is a factorization as in part (i), and the uniqueness assertion there gives $c(fg)$ an associate of $c(f)c(g)$ and $(fg)^*$ an associate of f^*g^*.

(iii) If $q \in R$, then it is obvious that $f(x) = qg^*(x) \in R[x]$. Conversely, if $f(x) \in R[x]$, then there is no need to clear denominators, and so $c(f) = d \in R$, where d is the gcd of the coefficients of $f(x)$. Thus, $f(x) = df^*(x)$. By uniqueness, there is a unit $w \in R$ with $q = wd \in R$.

(iv) Since $bf = hg^*$, we have $bc(f)f^* = c(h)h^*g^* = c(h)(hg)^*$. By uniqueness, f^*, $(hg)^*$, and h^*g^* are associates, and so $g^* \mid f^*$. But $f = c(f)f^*$, and so $g^* \mid f$. •

Definition. Let R be a UFD with $Q = \mathrm{Frac}(R)$. If $f(x) \in Q[x]$, there is a factorization $f(x) = c(f)f^*(x)$, where $c(f) \in Q$ and $f^*(x) \in R[x]$ is primitive. We call $c(f)$ the ***content*** of $f(x)$ and $f^*(x)$ the ***associated primitive polynomial***.

In light of Lemma 6.24(i), both $c(f)$ and $f^*(x)$ are essentially unique, differing only by a unit in R.

Theorem 6.25 (*Gauss*). *If R is a* UFD, *then $R[x]$ is also a* UFD.

Proof. We show first, by induction on $\deg(f)$, that every $f(x) \in R[x]$, neither zero nor a unit, is a product of irreducibles. If $\deg(f) = 0$, then $f(x)$ is a constant, hence lies in R. Since R is a UFD, f is a product of irreducibles. If $\deg(f) > 0$, then $f(x) =$

$c(f)f^*(x)$, where $c(f) \in R$ and $f^*(x)$ is primitive. Now $c(f)$ is either a unit or a product of irreducibles, by the base step. If $f^*(x)$ is irreducible, we are done. Otherwise, $f^*(x) = g(x)h(x)$, where neither g nor h is a unit. Since $f^*(x)$ is primitive, however, neither g nor h is a constant; therefore, each of these has degree less than $\deg(f^*) = \deg(f)$, and so each is a product of irreducibles, by the inductive hypothesis.

Proposition 6.17 now applies: $R[x]$ is a UFD if $(p(x))$ is a prime ideal for every irreducible $p(x) \in R[x]$; that is, if $p \mid fg$, then $p \mid f$ or $p \mid g$. Let us assume that $p(x) \nmid f(x)$.

Case (i). Suppose that $\deg(p) = 0$. Write

$$f(x) = c(f)f^*(x) \text{ and } g(x) = c(g)g^*(x),$$

where $c(f), c(g) \in R$, and $f^*(x), g^*(x)$ are primitive. Now $p \mid fg$, so that

$$p \mid c(f)c(g)f^*(x)g^*(x).$$

Since $f^*(x)g^*(x)$ is primitive, Lemma 6.24(ii) says that $c(f)c(g)$ an associate of $c(fg)$. However, if $p \mid f(x)g(x)$, then p divides each coefficient of fg; that is, p is a common divisor of all the coefficients of fg, and hence in R, which is a UFD, p divides the associates $c(fg)$ and $c(f)c(g)$. But Proposition 6.17 says that (p) is a prime ideal in R, and so $p \mid c(f)$ or $p \mid c(g)$. If $p \mid c(f)$, then p divides $c(f)f^*(x) = f(x)$, a contradiction. Therefore, $p \mid c(g)$ and, hence, $p \mid g(x)$, as desired.

Case (ii). Suppose that $\deg(p) > 0$. Let

$$(p, f) = \{s(x)p(x) + t(x)f(x) : s(x), t(x) \in R[x]\};$$

of course, (p, f) is an ideal containing $p(x)$ and $f(x)$. Choose $m(x) \in (p, f)$ of minimal degree. If $Q = \text{Frac}(R)$ is the fraction field of R, then the division algorithm in $Q[x]$ gives polynomials $q'(x), r'(x) \in Q[x]$ with

$$f(x) = m(x)q'(x) + r'(x),$$

where either $r'(x) = 0$ or $\deg(r') < \deg(m)$. Clearing denominators, there are polynomials $q(x), r(x) \in R[x]$ and a constant $b \in R$ with

$$bf(x) = q(x)m(x) + r(x),$$

where $r(x) = 0$ or $\deg(r) < \deg(m)$. Since $m \in (p, f)$, there are polynomials $s(x), t(x) \in R[x]$ with $m(x) = s(x)p(x) + t(x)f(x)$; hence $r = bf - qm \in (p, f)$. Since m has minimal degree in (p, f), we must have $r = 0$; that is, $bf(x) = m(x)q(x)$, and so $bf(x) = c(m)m^*(x)q(x)$. But $m^*(x)$ is primitive, and $m^*(x) \mid bf(x)$, so that $m^*(x) \mid f(x)$, by Lemma 6.24(iv). A similar argument, replacing $f(x)$ by $p(x)$ (that is, beginning with an equation $b''p(x) = q''(x)m(x) + r''(x)$ for some constant b''), gives $m^*(x) \mid p(x)$. Since $p(x)$ is irreducible, its only factors are units and associates. If $m^*(x)$ were an associate of $p(x)$, then $p(x) \mid f(x)$ (because $p(x) \mid m^*(x) \mid f(x)$), contrary to the hypothesis. Hence,

$m^*(x)$ must be a unit; that is, $m(x) = c(m) \in R$, and so (p, f) contains the nonzero constant $c(m)$. Now $c(m) = sp + tf$, and so

$$c(m)g(x) = s(x)p(x)g(x) + t(x)f(x)g(x).$$

Since $p(x) \mid f(x)g(x)$, we have $p(x) \mid c(m)g(x)$. But $p(x)$ is primitive, because it is irreducible, by Example 6.22, and so Lemma 6.24(iv) gives $p(x) \mid g(x)$. •

Corollary 6.26. *If k is a field, then $k[x_1, \ldots, x_n]$ is a* UFD.

Proof. The proof is by induction on $n \geq 1$. We proved, in Chapter 3, that the polynomial ring $k[x_1]$ in one variable is a UFD. For the inductive step, recall that $k[x_1, \ldots, x_n, x_{n+1}] = R[x_{n+1}]$, where $R = k[x_1, \ldots, x_n]$. By induction, R is a UFD, and by Corollary 6.25, so is $R[x_{n+1}]$. •

Proposition 6.20 shows that if k is a field, then gcd's exist in $k[x_1, \ldots, x_n]$.

Corollary 6.27 (Gauss). *Let R be a* UFD, *let $Q = \mathrm{Frac}(R)$, and let $f(x) \in R[x]$. If*

$$f(x) = G(x)H(x) \text{ in } Q[x],$$

then there is a factorization

$$f(x) = g(x)h(x) \text{ in } R[x],$$

where $\deg(g) = \deg(G)$ and $\deg(h) = \deg(H)$; in fact, $G(x)$ is a constant multiple of $g(x)$ and $H(x)$ is a constant multiple of $h(x)$. Therefore, if $f(x)$ does not factor into polynomials of smaller degree in $R[x]$, then $f(x)$ is irreducible in $Q[x]$.

Proof. By Lemma 6.24(i), the factorization $f(x) = G(x)H(x)$ in $Q[x]$ gives $q, q' \in Q$ with

$$f(x) = qG^*(x)q'H^*(x) \text{ in } Q[x],$$

where $G^*(x), H^*(x) \in R[x]$ are primitive. But $G^*(x)H^*(x)$ is primitive, by Gauss's lemma. Since $f(x) \in R[x]$, Lemma 6.24(iii) applies to say that the equation $f(x) = qq'[G^*(x)H^*(x)]$ forces $qq' \in R$. Therefore, $qq'G^*(x) \in R[x]$, and a factorization of $f(x)$ in $R[x]$ is $f(x) = [qq'G^*(x)]H^*(x)$. •

The special case $R = \mathbb{Z}$ and $Q = \mathbb{Q}$ is, of course, important.

Example 6.28.
We claim that $f(x, y) = x^2 + y^2 - 1 \in k[x, y]$ is irreducible, where k is a field. Write $Q = k(y) = \mathrm{Frac}(k[y])$, and view $f(x, y) \in Q[x]$. Now the quadratic $g(x) = x^2 + (y^2 - 1)$ is irreducible in $Q[x]$ if and only if it has no roots in $Q = k(y)$, and this is so, by Exercise 3.34 on page 142.

It follows from Proposition 6.17 that $(x^2 + y^2 - 1)$ is a prime ideal because it is generated by an irreducible polynomial. ◂

Recall that a complex number is an *algebraic integer* if it is a root of a monic polynomial in $\mathbb{Z}[x]$. Each algebraic integer has an irreducible polynomial associated with it.

Corollary 6.29. *If α is an algebraic integer, then* $\mathrm{irr}(\alpha, \mathbb{Q})$ *lies in* $\mathbb{Z}[x]$.

Proof. Let $p(x) \in \mathbb{Z}[x]$ be the monic polynomial of least degree having α as a root. If $p(x) = G(x)H(x)$ in $\mathbb{Q}[x]$, where $\deg(G) < \deg(p)$ and $\deg(H) < \deg(p)$, then α is a root of either $G(x)$ or $H(x)$. By Gauss's theorem, there is a factorization $p(x) = g(x)h(x)$ in $\mathbb{Z}[x]$ with $\deg(g) = \deg(G)$ and $\deg(h) = \deg(H)$; in fact, there are rationals c and d with $g(x) = cG(x)$ and $h(x) = dH(x)$. If a is the leading coefficient of $g(x)$ and b is the leading coefficient of $h(x)$, then $ab = 1$, for $p(x)$ is monic. Therefore, we may assume that $a = 1 = b$, for $a, b \in \mathbb{Z}$; that is, we may assume that both $g(x)$ and $h(x)$ are monic. Since α is a root of $g(x)$ or $h(x)$, we have contradicted $p(x)$ being a monic polynomial in $\mathbb{Z}[x]$ of least degree having α as a root. It follows that $p(x) = \mathrm{irr}(\alpha, \mathbb{Q})$, for the latter is the unique monic irreducible polynomial in $\mathbb{Q}[x]$ having α as a root. •

Definition. If α is an algebraic integer, then its ***minimal polynomial*** is the monic polynomial in $\mathbb{Z}[x]$ of least degree having α as a root.

Corollary 6.29 shows that every algebraic integer α has a unique minimal polynomial $m(x) \in \mathbb{Z}[x]$, namely, $m(x) = \mathrm{irr}(\alpha, \mathbb{Q})$, and $m(x)$ is irreducible in $\mathbb{Q}[x]$.

Remark. We define the (algebraic) ***conjugates*** of α to be the roots of $\mathrm{irr}(\alpha, \mathbb{Q})$, and we define the ***norm*** of α to be the absolute value of the product of the conjugates of α. Of course, the norm of α is just the absolute value of the constant term of $\mathrm{irr}(\alpha, \mathbb{Q})$, and so it is an (ordinary) integer. Norms are very useful in algebraic number theory, as we have seen in the proof of Theorem 3.66: Fermat's two-squares theorem. We have also considered them in the proof of Hilbert's Theorem 90, which was used to prove that if the Galois group of a polynomial $f(x) \in k[x]$ is solvable, where k has characteristic 0, then $f(x)$ is solvable by radicals. ◂

The next criterion uses the integers mod p.

Theorem 6.30. *Let* $f(x) = a_0 + a_1x + a_2x^2 + \cdots + x^n \in \mathbb{Z}[x]$ *be monic, and let p be a prime. If $f(x)$ is irreducible* mod p*, that is, if*

$$\widetilde{f}(x) = [a_0] + [a_1]x + [a_2]x^2 + \cdots + x^n \in \mathbb{F}_p[x],$$

is irreducible, then $f(x)$ is irreducible in $\mathbb{Q}[x]$.

Proof. By Proposition 3.48, the natural map $\varphi\colon \mathbb{Z} \to \mathbb{F}_p$ defines a homomorphism $\widetilde{\varphi}\colon \mathbb{Z}[x] \to \mathbb{F}_p[x]$ by

$$\widetilde{\varphi}(b_0 + b_1x + b_2x^2 + \cdots) = [b_0] + [b_1]x + [b_2]x^2 + \cdots;$$

that is, just reduce all the coefficients mod p. If $g(x) \in \mathbb{Z}[x]$, denote its image $\widetilde{\varphi}(g(x)) \in \mathbb{F}_p[x]$ by $\widetilde{g}(x)$. Suppose that $f(x)$ factors in $\mathbb{Z}[x]$; say, $f(x) = g(x)h(x)$, where $\deg(g) < \deg(f)$ and $\deg(h) < \deg(f)$; of course, $\deg(f) = \deg(g) + \deg(h)$. Now $\widetilde{f}(x) = \widetilde{g}(x)\widetilde{h}(x)$, because $\widetilde{\varphi}$ is a ring homomorphism, so that $\deg(\widetilde{f}) = \deg(\widetilde{g}) + \deg(\widetilde{h})$. Since

$f(x)$ is monic, $\widetilde{f}(x)$ is also monic, and so $\deg(\widetilde{f}) = \deg(f)$. Thus, both $\widetilde{g}(x)$ and $\widetilde{h}(x)$ have degrees less than $\deg(\widetilde{f})$, contradicting the irreducibility of $\widetilde{f}(x)$. Therefore, $f(x)$ is irreducible in $\mathbb{Z}[x]$, and, by Gauss's theorem, Corollary 6.27, $f(x)$ is irreducible in $\mathbb{Q}[x]$. •

Example 6.31.
The converse of Theorem 6.30 is false. It is easy to find an irreducible polynomial $f(x) \in \mathbb{Z}[x] \subseteq \mathbb{Q}[x]$ with $f(x)$ factoring mod p for some prime p, but we now show that $f(x) = x^4 + 1$ is an irreducible polynomial in $\mathbb{Z}[x]$ that factors mod p for every prime p.

First, $f(x)$ is irreducible in $\mathbb{Q}[x]$. By Corollary 3.44, $f(x)$ has no rational roots, and so the only possible factorization in $\mathbb{Q}[x]$ has the form

$$x^4 + 1 = (x^2 + ax + b)(x^2 - ax + c).$$

Multiplying out and equating like coefficients gives

$$\begin{aligned} c + b - a^2 &= 0 \\ a(c - b) &= 0 \\ bc &= 1. \end{aligned}$$

The second equation forces $a = 0$ or $c = b$, and it is quickly seen that either possibility leads to a contradiction.

We now show, for all primes p, that $x^4 + 1$ is not irreducible in $\mathbb{F}_p[x]$. If $p = 2$, then $x^4 + 1 = (x + 1)^4$, and so we may assume that p is an odd prime. As we saw in Example 1.21(i), every square is congruent to 0, 1, or 4 mod 8; since p is odd, we must have $p^2 \equiv 1 \bmod 8$. Therefore, $|(\mathbb{F}_{p^2})^\times| = p^2 - 1$ is divisible by 8. But $(\mathbb{F}_{p^2})^\times$ is a cyclic group, and so it has a (cyclic) subgroup of order 8, by Lemma 2.85. It follows that $\mathbb{F}_{p^2}$ contains all the 8th roots of unity; in particular, $\mathbb{F}_{p^2}$ contains all the roots of $x^4 + 1$. Hence, the splitting field E_p of $x^4 + 1$ over $\mathbb{F}_p$ is $\mathbb{F}_{p^2}$, and so $[E_p : \mathbb{F}_p] = 2$. But if $x^4 + 1$ were irreducible in $\mathbb{F}_p[x]$, then $4 \mid [E_p : \mathbb{F}_p]$, by Corollary 4.9. Therefore, $x^4 + 1$ factors in $\mathbb{F}_p[x]$ for every prime p. ◂

Theorem 6.30 says that if we can find a prime p with $\widetilde{f}(x)$ irreducible in $\mathbb{F}_p[x]$, then $f(x)$ is irreducible in $\mathbb{Q}[x]$. Until now, the finite fields $\mathbb{F}_p$ have been oddities; $\mathbb{F}_p$ has appeared only as a curious artificial construct. Now the finiteness of $\mathbb{F}_p$ is a genuine advantage, for there are only a finite number of polynomials in $\mathbb{F}_p[x]$ of any given degree. In Examples 3.35(i) and 3.35(ii), we displayed all the monic irreducible polynomials over $\mathbb{F}_2$ and over $\mathbb{F}_3$ of degree ≤ 3. In principle, then, we can test whether a polynomial of degree n in $\mathbb{F}_p[x]$ is irreducible by just looking at *all* the possible factorizations of it.

Example 6.32.
(i) We show that $f(x) = x^4 - 5x^3 + 2x + 3$ is an irreducible polynomial in $\mathbb{Q}[x]$.

By Corollary 3.44, the only candidates for rational roots of $f(x)$ are 1, −1, 3, −3, and the reader may check that none of these is a root. Since $f(x)$ is a quartic, we cannot yet conclude that $f(x)$ is irreducible, for it might be a product of (irreducible) quadratics.

Let us try the criterion of Theorem 6.30. Since $\widetilde{f}(x) = x^4 + x^3 + 1$ in $\mathbb{F}_2[x]$ is irreducible, by Example 3.35(i), it follows that $f(x)$ is irreducible in $\mathbb{Q}[x]$. [It was not necessary to check that $f(x)$ has no rational roots; the irreducibility of $\widetilde{f}(x)$ is enough to conclude the irreducibility of $f(x)$.]

(ii) Let $\Phi_5(x) = x^4 + x^3 + x^2 + x + 1 \in \mathbb{Q}[x]$.

In Example 3.35(i), we saw that $\widetilde{\Phi}_5(x) = x^4 + x^3 + x^2 + x + 1$ is irreducible in $\mathbb{F}_2[x]$, and so $\Phi_5(x)$ is irreducible in $\mathbb{Q}[x]$. ◂

Recall that if n is a positive integer, then the nth *cyclotomic polynomial* is

$$\Phi_n(x) = \prod (x - \zeta),$$

where ζ ranges over all the primitive nth roots of unity.

By Proposition 1.37, for every integer $n \geq 1$,

$$x^n - 1 = \prod_{d|n} \Phi_d(x),$$

where d ranges over all the divisors d of n. Now $\Phi_1(x) = x - 1$. When p is prime, then

$$x^p - 1 = \Phi_1(x)\Phi_p(x) = (x - 1)\Phi_p(x),$$

and so

$$\Phi_p(x) = (x^p - 1)/(x - 1) = x^{p-1} + x^{p-2} + \cdots + x + 1.$$

As any linear polynomial, $\Phi_2(x) = x + 1$ is irreducible in $\mathbb{Q}[x]$; the cyclotomic polynomial $\Phi_3(x) = x^2 + x + 1$ is irreducible in $\mathbb{Q}[x]$ because it has no rational roots; we have just seen that $\Phi_5(x)$ is irreducible in $\mathbb{Q}[x]$. Let us introduce another irreducibility criterion in order to prove that $\Phi_p(x)$ is irreducible in $\mathbb{Q}[x]$ for all primes p.

Lemma 6.33. *Let $g(x) \in \mathbb{Z}[x]$. If there is $c \in \mathbb{Z}$ with $g(x + c)$ irreducible in $\mathbb{Z}[x]$, then $g(x)$ is irreducible in $\mathbb{Q}[x]$.*

Proof. By Exercise 3.43 on page 149, the function $\varphi \colon \mathbb{Z}[x] \to \mathbb{Z}[x]$, given by $f(x) \mapsto f(x + c)$, is an isomorphism. If $g(x) = s(x)t(x)$, then $g(x + c) = \varphi(g(x)) = \varphi(st) = \varphi(s)\varphi(t)$ is a forbidden factorization of $g(x + c)$. Therefore, $g(x)$ is irreducible in $\mathbb{Z}[x]$ and hence, by Gauss's theorem, Corollary 6.27, $g(x)$ is irreducible in $\mathbb{Q}[x]$. •

The next result was found by G. Eisenstein. The following elegant proof of Eisenstein's criterion is in a 1969 paper of R. Singer; see Montgomery–Ralston, *Selected Papers in Algebra*.

Theorem 6.34 (Eisenstein Criterion). *Let R be a* UFD *with $Q = \operatorname{Frac}(R)$, and let $f(x) = a_0 + a_1x + \cdots + a_nx^n \in R[x]$. If there is an irreducible element $p \in R$ with $p \mid a_i$ for all $i < n$ but with $p \nmid a_n$ and $p^2 \nmid a_0$, then $f(x)$ is irreducible in $Q[x]$.*

Proof. Let $\widetilde{\varphi}\colon \mathbb{Z}[x] \to \mathbb{F}_p[x]$ be the ring homomorphism that reduces coefficients mod p, and let $\widetilde{f}(x)$ denote $\widetilde{\varphi}(f(x))$. If $f(x)$ is not irreducible in $\mathbb{Q}[x]$, then Gauss's theorem, Corollary 6.27, gives polynomials $g(x), h(x) \in \mathbb{Z}[x]$ with $f(x) = g(x)h(x)$, where $g(x) = b_0 + b_1 x + \cdots + b_m x^m$ and $h(x) = c_0 + c_1 x + \cdots + c_k x^k$. There is thus an equation $\widetilde{f}(x) = \widetilde{g}(x)\widetilde{h}(x)$ in $\mathbb{F}_p[x]$.

Since p does not divide a_n, we have $\widetilde{f}(x) \neq 0$; in fact, $\widetilde{f}(x) = ux^n$ for some unit $u \in \mathbb{F}_p$, because all its coefficients aside from its leading coefficient are 0. By Theorem 3.42, unique factorization in $\mathbb{F}_p[x]$, we must have $\widetilde{g}(x) = vx^m$, where v is a unit in $\mathbb{F}_p$, for any irreducible factor of $\widetilde{g}(x)$ is an irreducible factor of $\widetilde{f}(x)$; similarly, $\widetilde{h}(x) = wx^k$, where w is a unit in $\mathbb{F}_p$. It follows that each of $\widetilde{g}(x)$ and $\widetilde{h}(x)$ has constant term 0; that is, $[b_0] = 0 = [c_0]$ in $\mathbb{F}_p$; equivalently, $p \mid b_0$ and $p \mid c_0$. But $a_0 = b_0 c_0$, and so $p^2 \mid a_0$, a contradiction. Therefore, $f(x)$ is irreducible in $\mathbb{Q}[x]$. •

Of course, Eisenstein's criterion holds for polynomials in $\mathbb{Z}[x]$. The generalization from $\mathbb{Z}$ to PIDs is instantaneous.

Corollary 6.35 (*Gauss*). *For every prime p, the pth cyclotomic polynomial $\Phi_p(x)$ is irreducible in $\mathbb{Q}[x]$.*

Proof. Since $\Phi_p(x) = (x^p - 1)/(x - 1)$, we have

$$\begin{aligned}\Phi_p(x+1) &= [(x+1)^p - 1]/x \\ &= x^{p-1} + \binom{p}{1} x^{p-2} + \binom{p}{2} x^{p-3} + \cdots + p.\end{aligned}$$

Since p is prime, Proposition 1.12 shows that Eisenstein's criterion applies; we conclude that $\Phi_p(x+1)$ is irreducible in $\mathbb{Q}[x]$. By Lemma 6.33, $\Phi_p(x)$ is irreducible in $\mathbb{Q}[x]$. •

We do not say that $x^{n-1} + x^{n-2} + \cdots + x + 1$ is irreducible when n is not prime. For example, $x^3 + x^2 + x + 1 = (x+1)(x^2+1)$.

Irreducibility of a polynomial in several variables is more difficult to determine than irreducibility of a polynomial of one variable, but here is one criterion.

Proposition 6.36. *Let k be a field and let $f(x_1, \ldots, x_n)$ be a primitive polynomial in $R[x_n]$, where $R = k[x_1, \ldots, x_{n-1}]$. If f cannot be factored into two polynomials of lower degree in $R[x_n]$, then f is irreducible in $k[x_1, \ldots, x_n]$.*

Proof. Let us write $f(x_1, \ldots, x_n) = F(x_n)$ if we wish to view f as a polynomial in $R[x_n]$ (of course, the coefficients of F are polynomials in $k[x_1, \ldots, x_{n-1}]$). Suppose that $F(x_n) = G(x_n)H(x_n)$; by hypothesis, the degrees of G and H (in x_n) cannot both be less than $\deg(F)$, and so one of them, say, G, has degree 0. It follows, because F is primitive, that G is a unit in $k[x_1, \ldots, x_{n-1}]$. Therefore, $f(x_1, \ldots, x_n)$ is irreducible in $R[x_n] = k[x_1, \ldots, x_n]$. •

Of course, the proposition applies to any variable x_i, not just to x_n.

Corollary 6.37. *If k is a field and $g(x_1, \ldots, x_n), h(x_1, \ldots, x_n) \in k[x_1, \ldots, x_n]$ are relatively prime, then $f(x_1, \ldots, x_n, y) = yg(x_1, \ldots, x_n) + h(x_1, \ldots, x_n)$ is irreducible in $k[x_1, \ldots, x_n, y]$.*

Proof. Let $R = k[x_1, \cdots, x_n]$. Note that f is primitive in $R[y]$, because $(g, h) = 1$ forces any divisor of its coefficients g, h to be a unit. Since f is linear in y, it is not the product of two polynomials in $R[y]$ of smaller degree, and hence Proposition 6.36 shows that f is irreducible in $R[y] = k[x_1, \ldots, x_n, y]$. •

For example, $xy^2 + z$ is an irreducible polynomial in $k[x, y, z]$ because it is a primitive polynomial that is linear in x.

EXERCISES

6.17 Let R be a UFD and let $Q = \text{Frac}(R)$ be its fraction field. Prove that each nonzero $a/b \in Q$ has an expression in lowest terms; that is, a and b are relatively prime.

6.18 Let R be a UFD.

(i) If $a, b, c \in R$ and a and b are relatively prime, prove that $a \mid bc$ implies $a \mid c$.

(ii) If $a, c_1, \ldots, c_n \in R$ and $c_i \mid a$ for all i, prove that $c \mid a$, where $c = \text{lcm}\{c_1, \ldots, c_n\}$.

6.19 If R is a domain, prove that the only units in $R[x_1, \ldots, x_n]$ are units in R. On the other hand, prove that $2x + 1$ is a unit in $\mathbb{I}_4[x]$.

6.20 Prove that a UFD R is a PID if and only if every nonzero prime ideal is a maximal ideal.

6.21 (i) Prove that x and y are relatively prime in $k[x, y]$, where k is a field.

(ii) Prove that 1 is not a linear combination of x and y in $k[x, y]$.

6.22 (i) Prove that $\mathbb{Z}[x_1, \ldots, x_n]$ is a UFD for all $n \geq 1$.

(ii) If R is a field, prove that the ring of polynomials in infinitely many variables, $R = k[x_1, x_2, \ldots, x_n \ldots]$, is also a UFD.

Hint. We have not given a formal definition of R (it will be given in Chapter 8), but, for the purposes of this exercise, regard R as the union of the ascending chain $k[x_1] \subsetneq k[x_1, x_2] \subsetneq \cdots \subsetneq k[x_1, x_2, \ldots, x_n] \subsetneq \cdots$.

6.23 Determine whether the following polynomials are irreducible in $\mathbb{Q}[x]$.

(i) $f(x) = 3x^2 - 7x - 5$.

(ii) $f(x) = 2x^3 - x - 6$.

(iii) $f(x) = 8x^3 - 6x - 1$.

(iv) $f(x) = x^3 + 6x^2 + 5x + 25$.

(v) $f(x) = x^4 + 8x + 12$.

Hint. In $\mathbb{F}_5[x]$, $f(x) = (x + 1)g(x)$, where $g(x)$ is irreducible.

(vi) $f(x) = x^5 - 4x + 2$.

(vii) $f(x) = x^4 + x^2 + x + 1$.

Hint. Show that $f(x)$ has no roots in $\mathbb{F}_3$ and that a factorization of $f(x)$ as a product of quadratics would force impossible restrictions on the coefficients.

(viii) $f(x) = x^4 - 10x^2 + 1$.

Hint. Show that $f(x)$ has no rational roots and that a factorization of $f(x)$ as a product of quadratics would force impossible restrictions on the coefficients.

6.24 Is $x^5 + x + 1$ irreducible in $\mathbb{F}_2[x]$?
Hint. Use Example 3.35(i).

6.25 Let $f(x) = (x^p - 1)/(x - 1)$, where p is prime. Using the identity

$$f(x+1) = x^{p-1} + pq(x),$$

where $q(x) \in \mathbb{Z}[x]$ has constant term 1, prove that $x^{p^n(p-1)} + \cdots + x^{p^n} + 1$ is irreducible in $\mathbb{Q}[x]$ for all $n \geq 0$.

6.26 (i) If a is a squarefree integer, prove that $x^n - a$ is irreducible in $\mathbb{Q}[x]$ for every $n \geq 1$. Conclude that there are irreducible polynomials in $\mathbb{Q}[x]$ of every degree $n \geq 1$.
Hint. Use the Eisenstein criterion.

(ii) If a is a squarefree integer, prove that $\sqrt[n]{a}$ is irrational for all $n \geq 2$.

6.27 Let k be a field, and let $f(x) = a_0 + a_1 x + \cdots + a_n x^n \in k[x]$ have degree n and nonzero constant term a_0. Prove that if $f(x)$ is irreducible, then so is $a_n + a_{n-1}x + \cdots + a_0 x^n$.

6.28 Let k be a field and let $f(x_1, \ldots, x_n) \in k[x_1, \ldots, x_n]$ be a primitive polynomial in $R[x_n]$, where $R = k[x_1, \ldots, x_{n-1}]$. If f is either quadratic or cubic in x_n, prove that f is irreducible in $k[x_1, \ldots, x_n]$ if and only if f has no roots in $k(x_1, \ldots, x_{n-1})$.

6.29 Let R be a UFD with $Q = \mathrm{Frac}(R)$. If $f(x) \in R[x]$, prove that $f(x)$ is irreducible in $R[x]$ if and only if $f(x)$ is primitive and $f(x)$ is irreducible in $Q[x]$.

6.30 Prove that

$$f(x, y) = xy^3 + x^2y^2 - x^5y + x^2 + 1$$

is an irreducible polynomial in $\mathbb{R}[x, y]$.

6.31 Let $D = \det\left(\begin{bmatrix} x & y \\ z & w \end{bmatrix}\right)$, so that D lies in the polynomial ring $\mathbb{Z}[x, y, z, w]$.

(i) Prove that (D) is a prime ideal in $\mathbb{Z}[x, y, z, w]$.
Hint. Prove first that D is an irreducible element.

(ii) Prove that $\mathbb{Z}[x, y, z, w]/(D)$ is not a UFD. Another example of a domain which is not a UFD is given in Example 6.21.

6.3 Noetherian Rings

One of the most important properties of $k[x_1, \ldots, x_n]$, when k is a field, is that every ideal in it can be generated by a finite number of elements. This property is intimately related to chains of ideals, which we have already seen in the course of proving that PIDs are UFDs (I apologize for so many acronyms, but here comes another one!).

Definition. A commutative ring R satisfies the **ACC**, the ***ascending chain condition***, if every ascending chain of ideals

$$I_1 \subseteq I_2 \subseteq \cdots \subseteq I_n \subseteq \cdots$$

stops; that is, the sequence is constant from some point on: there is an integer N with $I_N = I_{N+1} = I_{N+2} = \cdots$.

Lemma 6.18(ii) shows that every PID satisfies the ACC.
Here is an important type of ideal.

Definition. If X is a subset of a commutative ring R, then the ***ideal generated by*** X is the set of all finite linear combinations

$$I = (X) = \Big\{ \sum_{\text{finite}} r_i a_i : r_i \in R \text{ and } x_i \in X \Big\}.$$

We say that I is ***finitely generated***, often abbreviated to f.g., if $X = \{a_1, \ldots, a_n\}$; that is, every element in I is an R-linear combination of the a_i. We write

$$I = (a_1, \ldots, a_n),$$

and we call I the ***ideal generated by*** $a_1, \ldots, a_n$.

A set of generators $a_1, \ldots, a_n$ of an ideal I is sometimes called a ***basis*** of I (even though this is a weaker notion than that of a basis of a vector space, for we do not assume that the coefficients r_i in the expression $c = \sum r_i a_i$ are uniquely determined by c).

Of course, every ideal I in a PID is finitely generated, for it can be generated by one element.

Proposition 6.38. *The following conditions are equivalent for a commutative ring R.*

(i) *R has the* ACC.

(ii) *R satisfies the **maximum condition**: Every nonempty family $\mathcal{F}$ of ideals in R has a maximal element; that is, there is some $I_0 \in \mathcal{F}$ for which there is no $I \in \mathcal{F}$ with $I_0 \subsetneq I$.*

(iii) *Every ideal in R is finitely generated.*

Proof. (i) $\Rightarrow$ (ii): Let $\mathcal{F}$ be a family of ideals in R, and assume that $\mathcal{F}$ has no maximal element. Choose $I_1 \in \mathcal{F}$. Since I_1 is not a maximal element, there is $I_2 \in \mathcal{F}$ with $I_1 \subsetneq I_2$. Now I_2 is not a maximal element in $\mathcal{F}$, and so there is $I_3 \in \mathcal{F}$ with $I_2 \subsetneq I_3$. Continuing in this way, we can construct an ascending chain of ideals in R that does not stop, contradicting the ACC.

(ii) $\Rightarrow$ (iii): Let I be an ideal in R, and define $\mathcal{F}$ to be the family of all the finitely generated ideals contained in I; of course, $\mathcal{F} \neq \varnothing$ (for $\{0\} \in \mathcal{F}$). By hypothesis, there exists a maximal element $M \in \mathcal{F}$. Now $M \subseteq I$ because $M \in \mathcal{F}$. If $M \subsetneq I$, then there is $a \in I$ with $a \notin M$. The ideal

$$J = \{m + ra : m \in M \text{ and } r \in R\} \subseteq I$$

is finitely generated, and so $J \in \mathcal{F}$; but $M \subsetneq J$, and this contradicts the maximality of M. Therefore, $M = I$, and so I is finitely generated.

(iii) $\Rightarrow$ (i): Assume that every ideal in R is finitely generated, and let

$$I_1 \subseteq I_2 \subseteq \cdots \subseteq I_n \subseteq \cdots$$

be an ascending chain of ideals in R. By Lemma 6.18(i), the ascending union $J = \bigcup_{n \geq 1} I_n$ is an ideal.

By hypothesis, there are elements $a_i \in J$ with $J = (a_1, \ldots, a_q)$. Now a_i got into J by being in I_{n_i} for some n_i. If N is the largest n_i, then $I_{n_i} \subseteq I_N$ for all i; hence, $a_i \in I_N$ for all i, and so

$$J = (a_1, \ldots, a_q) \subseteq I_N \subseteq J.$$

It follows that if $n \geq N$, then $J = I_N \subseteq I_n \subseteq J$, so that $I_n = J$; therefore, the chain stops, and R has the ACC. •

We now give a name to a commutative ring that satisfies any of the three equivalent conditions in the proposition.

Definition. A commutative ring R is called ***noetherian***[4] if every ideal in R is finitely generated.

We shall soon see that $k[x_1, \ldots, x_n]$ is noetherian whenever k is a field. On the other hand, here is an example of a commutative ring that is not noetherian.

Example 6.39.
Let $R = \mathcal{F}(\mathbb{R})$ be the ring of all real-valued functions on the reals, under pointwise operations (see Example 3.7). It is easy to see, for every positive integer n, that

$$I_n = \{f : \mathbb{R} \to \mathbb{R} : f(x) = 0 \text{ for all } x \geq n\}$$

is an ideal and that $I_n \subsetneq I_{n+1}$ for all n. Therefore, R does not satisfy the ACC, and so R is not noetherian. ◄

Here is an application of the maximum condition.

Corollary 6.40. *If I is a proper ideal in a noetherian ring R, then there exists a maximal ideal M in R containing I. In particular, every noetherian ring has maximal ideals.*[5]

Proof. Let $\mathcal{F}$ be the family of all those proper ideals in R which contain I; note that $\mathcal{F} \neq \varnothing$ because $I \in \mathcal{F}$. Since R is noetherian, the maximum condition gives a maximal element M in $\mathcal{F}$. We must still show that M is a maximal ideal in R (that is, that M is a maximal element in the larger family $\mathcal{F}'$ consisting of all the proper ideals in R). Suppose there is a proper ideal J with $M \subseteq J$. Then $I \subseteq J$, and so $J \in \mathcal{F}$; therefore, maximality of M gives $M = J$, and so M is a maximal ideal in R. •

Here is one way to construct a new noetherian ring from an old one.

[4]This name honors Emmy Noether (1882–1935), who introduced chain conditions in 1921.

[5]This corollary is true without assuming R is noetherian, but the proof of the general result needs Zorn's lemma; see Theorem 6.46.

Corollary 6.41. *If R is a noetherian ring and I is an ideal in R, then R/I is also noetherian.*

Proof. If A is an ideal in R/I, then the correspondence theorem provides an ideal J in R with $J/I = A$. Since R is noetherian, the ideal J is finitely generated, say, $J = (b_1, \ldots, b_n)$, and so $A = J/I$ is also finitely generated (by the cosets $b_1 + I, \ldots, b_n + I$). Therefore, R/I is noetherian. •

The following anecdote is well known. Around 1890, Hilbert proved the famous Hilbert basis theorem, showing that every ideal in $\mathbb{C}[x_1, \ldots, x_n]$ is finitely generated. As we will see, the proof is nonconstructive in the sense that it does not give an explicit set of generators of an ideal. It is reported that when P. Gordan, one of the leading algebraists of the time, first saw Hilbert's proof, he said, "This is not mathematics, but theology!" On the other hand, Gordan said, in 1899 when he published a simplified proof of Hilbert's theorem, "I have convinced myself that theology also has its advantages."

The proof of the Hilbert basis theorem given next is due to H. Sarges (1976).

Theorem 6.42 (Hilbert Basis Theorem). *If R is a commutative noetherian ring, then $R[x]$ is also noetherian.*

Proof. Assume that I is an ideal in $R[x]$ that is not finitely generated; of course, $I \neq \{0\}$. Define $f_0(x)$ to be a polynomial in I of minimal degree and define, inductively, $f_{n+1}(x)$ to be a polynomial of minimal degree in $I - (f_0, \ldots, f_n)$. It is clear that

$$\deg(f_0) \leq \deg(f_1) \leq \deg(f_2) \leq \cdots.$$

Let a_n denote the leading coefficient of $f_n(x)$. Since R is noetherian, Exercise 6.32 on page 344 applies to give an integer m with $a_{m+1} \in (a_0, \ldots, a_m)$; that is, there are $r_i \in R$ with $a_{m+1} = r_0 a_0 + \cdots + r_m a_m$. Define

$$f^*(x) = f_{m+1}(x) - \sum_{i=0}^{m} x^{d_{m+1}-d_i} r_i f_i(x),$$

where $d_i = \deg(f_i)$. Now $f^*(x) \in I - (f_0(x), \ldots, f_m(x))$, otherwise $f_{m+1}(x) \in (f_0(x), \ldots, f_m(x))$. It suffices to show that $\deg(f^*) < \deg(f_{m+1})$, for this contradicts $f_{m+1}(x)$ having minimal degree among polynomials in I that are not in $(f_0, \ldots, f_m)$. If $f_i(x) = a_i x^{d_i}$ + lower terms, then

$$\begin{aligned} f^*(x) &= f_{m+1}(x) - \sum_{i=0}^{m} x^{d_{m+1}-d_i} r_i f_i(x) \\ &= (a_{m+1}x^{d_{m+1}} + \text{lower terms}) - \sum_{i=0}^{m} x^{d_{m+1}-d_i} r_i (a_i x^{d_i} + \text{lower terms}). \end{aligned}$$

The leading term being subtracted is thus $\sum_{i=0}^{m} r_i a_i x^{d_{m+1}} = a_{m+1} x^{d_{m+1}}$. •

Corollary 6.43.

(i) *If k is a field, then $k[x_1, \ldots, x_n]$ is noetherian.*

(ii) *The ring $\mathbb{Z}[x_1, \ldots, x_n]$ is noetherian.*

(iii) *For any ideal I in $k[x_1, \ldots, x_n]$, where $k = \mathbb{Z}$ or k is a field, the quotient ring $k[x_1, \ldots, x_n]/I$ is noetherian.*

Proof. The proofs of the first two items are by induction on $n \geq 1$, using the theorem, while the proof of item (iii) follows from Corollary 6.41. •

EXERCISES

6.32 Let R be a commutative ring. Prove that R is noetherian if and only if, for every sequence $a_1, a_2, \ldots, a_n, \ldots$ of elements in R, there is an integer $m \geq 1$ with a_{m+1} an R-linear combination of its predecessors; that is, there are $r_1, \ldots, r_m \in R$ with $a_{m+1} = r_1 a_1 + \cdots + r_m a_m$.

6.33 (i) Give an example of a noetherian ring R containing a subring that is not noetherian.

(ii) Give an example of a commutative ring R containing proper ideals $I \subsetneq J \subsetneq R$ with J finitely generated but with I not finitely generated.

6.34 Let R be a noetherian domain such that every $a, b \in R$ has a gcd that is an R-linear combination of a and b. Prove that R is a PID. (The noetherian hypothesis is necessary, for there exist non-noetherian domains, called *Bézout rings*, in which every finitely generated ideal is principal.)

Hint. Use induction on the number of generators of an ideal.

6.35 Give a proof that every nonempty family $\mathcal{F}$ of ideals in a PID R has a maximal element without using Proposition 6.38.

6.36 Example 6.39 shows that $R = \mathcal{F}(\mathbb{R})$, the ring of all functions on $\mathbb{R}$ under pointwise operations, does not satisfy the ACC.

(i) Show that the family of ideals $\{I_n : n \geq 1\}$ in that example does not have a maximal element.

(ii) Define

$$f_n(x) = \begin{cases} 1 & \text{if } x < n \\ 0 & \text{if } x \geq n, \end{cases}$$

and define $J_n = (f_1, \ldots, f_n)$. Prove that $J^* = \bigcup_{n \geq 1} J_n$ is an ideal that is not finitely generated.

6.37 If R is a commutative ring, define the ring of formal power series in several variables inductively:

$$R[[x_1, \ldots, x_{n+1}]] = A[[x_{n+1}]],$$

where $A = R[[x_1, \ldots, x_n]]$.

Prove that if R is a noetherian ring, then $R[[x_1, \ldots, x_n]]$, is also a noetherian ring.

Hint. Use Exercise 3.54(i) on page 151 if $n = 1$; use the proof of the Hilbert basis theorem when $n \geq 1$, but replace the degree of a polynomial by the order of a power series (where the *order* of a nonzero power series $\sum c_i x^i$ is n if n is the smallest i with $c_i \neq 0$).

6.38 Let
$$S^2 = \{(a, b, c) \in \mathbb{R}^3 \colon a^2 + b^2 + c^2 = 1\}$$
be the unit sphere in $\mathbb{R}^3$, and let
$$I = \{ f(x, y, z) \in \mathbb{R}[x, y, z] \colon f(a, b, c) = 0 \text{ for all } (a, b, c) \in S^2\}.$$
Prove that I is a finitely generated ideal in $\mathbb{R}[x, y, z]$.

6.39 If R and S are noetherian, prove that their direct product $R \times S$ is also noetherian.

6.40 If R is a commutative ring that is also a vector space over a field k, then R is called a ***commutative k-algebra*** if
$$(\alpha u)v = \alpha(uv) = u(\alpha v)$$
for all $\alpha \in k$ and $u, v \in R$. Prove that every commutative k-algebra that is finite-dimensional over k is noetherian.

6.4 APPLICATIONS OF ZORN'S LEMMA

Dealing with infinite sets may require some appropriate tools of set theory.

Definition. If A is a set, let $\mathcal{P}(A)^\#$ denote the family of all its nonempty subsets. The ***axiom of choice*** states that if A is a nonempty set, then there exists a function $\beta : \mathcal{P}(A)^\# \to A$ with $\beta(S) \in S$ for every nonempty subset S of A. Such a function β is called a ***choice function***.

Informally, the axiom of choice is a harmless looking statement; it says that we can simultaneously choose one element from each nonempty subset of a set.

The axiom of choice is easy to accept, and it is one of the standard axioms of set theory. Indeed, the axiom of choice is equivalent to the statement that the cartesian product of nonempty sets is itself nonempty (see Proposition A.1 in the Appendix). However, the axiom is not convenient to use as it stands. There are various equivalent forms of it that are more useful, the most popular of which are *Zorn's lemma* and the *well-ordering principle*.

Recall that a set X is a *partially ordered set* if there is a relation $x \preceq y$ defined on X that is reflexive, antisymmetric, and transitive.

We introduce some definitions to enable us to state the well-ordering principle.

Definition. A partially ordered set X is ***well-ordered*** if every nonempty subset S of X contains a ***smallest element***; that is, there is $s_0 \in S$ with
$$s_0 \preceq s \text{ for all } s \in S.$$

The set of natural numbers $\mathbb{N}$ is well-ordered (this is precisely what the least integer axiom in Chapter 1 states), but the set $\mathbb{Z}$ of all integers is not well-ordered because $\mathbb{Z}$ itself is a subset having no smallest element.

Well-ordering principle. *Every set X has some well-ordering of its elements.*

If X happens to be a partially ordered set, then a well-ordering, whose existence is asserted by the well-ordering principle, may have nothing to do with the original partial ordering. For example, $\mathbb{Z}$ can be well-ordered:

$$0 \preceq 1 \preceq -1 \preceq 2 \preceq -2 \preceq \cdots.$$

We will be able to state Zorn's lemma after the following definitions.

Definition. An element m in a partially ordered set X is a ***maximal element*** if there is no $x \in X$ for which $m \prec x$; that is,

$$\text{if } m \preceq x, \text{ then } m = x.$$

Recall that an *upper bound* of a nonempty subset Y of a partially ordered set X is an element $x_0 \in X$, not necessarily in Y, with $y \preceq x_0$ for every $y \in Y$.

Example 6.44.
(i) A partially ordered set may have no maximal elements. For example, $\mathbb{R}$, with its usual ordering, has no maximal elements.

(ii) A partially ordered set may have many maximal elements. For example, if X is the partially ordered set of all the proper subsets of a set U, then a subset S is a maximal element if and only if $S = U - \{u\}$ for some $u \in U$; that is, S is the complement of a point.

(iii) If X is the family of all the proper ideals in a commutative ring R, partially ordered by inclusion, then a maximal element in X is a maximal ideal. ◂

Zorn's lemma gives a condition that guarantees the existence of maximal elements.

Definition. A partially ordered set X is a ***chain*** if, for all $x, y \in X$, either $x \preceq y$ or $y \preceq x$.

The set of real numbers $\mathbb{R}$ is a chain if one takes $x \preceq y$ to be the usual inequality $x \leq y$.

Zorn's lemma. *If X is a nonempty partially ordered set in which every chain has an upper bound in X, then X has a maximal element.*

Theorem. *The following statements are equivalent*:

(i) *Zorn's lemma.*

(ii) *The well-ordering principle.*

(iii) *The axiom of choice.*

Proof. See the Appendix. •

Henceforth, we shall assume, unashamedly, that all these statements are true, and we will use any of them whenever convenient.

The next proposition is frequently used when verifying that the hypothesis of Zorn's lemma does hold.

Proposition 6.45. *If C is a chain and $S = \{x_1, \ldots, x_n\} \subseteq C$, then there exists some x_i, for $1 \le i \le n$, with $x_j \preceq x_i$ for all $x_j \in S$.*

Proof. The proof is by induction on $n \ge 1$. The base step is trivially true. Let $S = \{x_1, \ldots, x_{n+1}\}$, and define $S' = \{x_1, \ldots, x_n\}$. The inductive hypothesis provides x_i, for $1 \le i \le n$, with $x_j \preceq x_i$ for all $x_j \in S'$. Since C is a chain, either $x_i \preceq x_{n+1}$ or $x_{n+1} \preceq x_i$. Either case provides a largest element of S. •

Here is our first application of Zorn's lemma.

Theorem 6.46. *If R is a nonzero commutative ring, then R has a maximal ideal. Indeed, every proper ideal I in R is contained in a maximal ideal.*

Proof. The second statement implies the first, for if R is a nonzero ring, then the ideal (0) is a proper ideal, and so there exists a maximal ideal in R containing it.

Let X be the family of all the proper ideals containing I (note that $X \neq \varnothing$ because $I \in X$), and partially order X by inclusion. It is easy to see that a maximal element of X is a maximal ideal in R: There is no proper ideal strictly containing it.

Let $\mathcal{C}$ be a chain of X; thus, given $I, J \in \mathcal{C}$, either $I \subseteq J$ or $J \subseteq I$. We claim that $I^* = \bigcup_{I \in \mathcal{C}} I$ is an upper bound of $\mathcal{C}$. Clearly, $I \subseteq I^*$ for all $I \in \mathcal{C}$, so that it remains to prove that I^* is a proper ideal. The argument that I^* is an ideal is, by now, familiar. Finally, we show that I^* is a proper ideal. If $I^* = R$, then $1 \in I^*$; now 1 got into I^* because $1 \in I$ for some $I \in \mathcal{C}$, and this contradicts I being a proper ideal.

We have verified that every chain of X has an upper bound. Hence, Zorn's lemma provides a maximal element, as desired. •

Remark. Theorem 6.46 would be false if the definition of ring R did not insist on R containing 1. An example of such a "ring without unit" is any additive abelian group G with multiplication defined by $ab = 0$ for all $a, b \in G$. The usual definition of *ideal* makes sense, and it is easy to see that the ideals in G are its subgroups. Thus, a maximal ideal I is just a maximal subgroup, which means that G/I has no proper subgroups, by the correspondence theorem. Thus, G/I is a simple abelian group; that is, G/I is a finite group of prime order. In particular, take $G = \mathbb{Q}$ as an additive abelian group, and equip it with the zero multiplication. The reader can show that $\mathbb{Q}$ has no nonzero finite quotient groups, so that it has no maximal subgroups. Therefore, this "ring without unit" has no maximal ideals. ◂

We emphasize the necessity of checking, when applying Zorn's lemma to a partially ordered set X, that X be nonempty. For example, a careless person might claim that Zorn's lemma can be used to prove that there is a maximal uncountable subset of $\mathbb{Z}$. Define X to be the set of all the uncountable subsets of $\mathbb{Z}$, and partially order X by inclusion. If C is a chain in X, then it is clear that the uncountable subset $S^* = \bigcup_{S \subseteq C} S$ is an upper bound of C, for $S \subseteq S^*$ for every $S \in C$. Therefore, Zorn's lemma provides a maximal element of X, which must be a maximal uncountable subset of $\mathbb{Z}$. The flaw, of course, is that $X = \varnothing$ (for every subset of a countable set is itself countable).

Here is our second application of Zorn's lemma. We begin by generalizing the usual definition of a basis of a vector space so that it applies to all, not necessarily finite-dimensional, vector spaces.

Definition. Let V be a vector space over some field k, and let $Y \subseteq V$ be an infinite subset.[6]

(i) Y is ***linearly independent*** if every finite subset of Y is linearly independent.

(ii) Y ***spans*** V if each $v \in V$ is a linear combination of finitely[7] many elements of Y. We write $V = \langle Y \rangle$ when V is spanned by Y.

(iii) A ***basis*** of a vector space V is a linearly independent subset that spans V.

Thus, an infinite subset $Y = \{y_i : i \in I\}$ is linearly independent if, whenever $\sum a_i y_i = 0$ (where only finitely many $a_i \neq 0$), then $a_i = 0$ for all i.

Example 6.47.
Let k be a field, and let $V = k[x]$ regarded as a vector space over k. We claim that

$$Y = \{1, x, x^2, \ldots, x^n, \ldots\}$$

is a basis of V. Now Y spans V, for any polynomial of degree d is a k-linear combination of $1, x, x^2, \ldots, x^d$. Also, Y is linearly independent, because there are no scalars $a_0, a_1, \ldots, a_n$, not all 0, with $\sum_{i=0}^{n} a_i x^i = 0$ (a polynomial is the zero polynomial precisely if all its coefficients are 0). Therefore, Y is a basis of V. ◄

Theorem 6.48. *Every vector space V over a field F has a basis. Indeed, every linearly independent subset B of V is contained in a basis of V; that is, there is a subset B' so that $B \cup B'$ is a basis of V.*

Proof. Note that the first statement follows from the second, for $B = \varnothing$ is a linearly independent subset contained in a basis.

Let X be the family of all the linearly independent subsets of V that contain B. The family X is nonempty, for $B \in X$. Partially order X by inclusion. We use Zorn's lemma to prove the existence of a maximal element in X. Let $\mathcal{B} = \{B_j : j \in J\}$ be a chain of X. Thus, each B_j is a linearly independent subset containing B and, for all $i, j \in J$, either $B_j \subseteq B_i$ or $B_i \subseteq B_j$. It follows from Proposition 6.45 that if $B_{j_1}, \ldots, B_{j_n}$ is any *finite* family of B_j's, then one contains all of the others.

Let $B^* = \bigcup_{j \in J} B_j$. Clearly, B^* contains B and $B_j \subseteq B^*$ for all $j \in J$. Thus, B^* is an upper bound of $\mathcal{B}$ if it belongs to X; that is, if B^* is a linearly independent subset of V. If B^* is not linearly independent, then it has a finite subset $y_{i_1}, \ldots, y_{i_m}$ that is linearly dependent. How did y_{i_k} get into B^*? Answer: $y_{i_k} \in B_{j_k}$ for some index j_k. Since there are only finitely many y_{i_k}, there exists B_{j_0} containing all the B_{i_k}; that is, $y_{i_1}, \ldots, y_{i_m} \in B_{j_0}$.

[6] When dealing with infinite bases, it is more convenient to work with subsets instead of lists.

[7] Only finite sums of elements in V are allowed. Without limits, convergence of infinite series does not make sense, and so a sum with infinitely many nonzero terms is not defined.

But B_{j_0} is linearly independent, by hypothesis, and this is a contradiction. Therefore, B^* is an upper bound of the simply ordered subset $\mathcal{B}$. We have verified that every chain of X has an upper bound. Hence, Zorn's lemma applies to say that there is a maximal element in X.

Let M be a maximal element in X. Since M is linearly independent, it suffices to show that it spans V (for then M is a basis of V containing B). If M does not span V, then there is $v_0 \in V$ with $v_0 \notin \langle M \rangle$, the subspace spanned by M. Consider the subset $M^* = M \cup \{v_0\}$. Clearly, $M \subsetneq M^*$. Now M^* is linearly independent: if $a_0 v_0 + \sum a_i y_i = 0$, where $y_i \in M$ and $a_0, a_i \in F$ are not all 0, then $a_0 \neq 0$ (otherwise the collection of y_i appearing in the equation would be linearly dependent, a contradiction). But if $a_0 \neq 0$, then $v_0 = -a_0^{-1} \sum a_i y_i$, contradicting $v_0 \notin \langle M \rangle$. Therefore, M is a basis of V. The last statement follows if we define $B' = M - B$. •

Recall that a subspace W of a vector space V is a *direct summand* if there is a subspace W' of V with $\{0\} = W \cap W'$ and $V = W + W'$ (i.e., each $v \in V$ can be written $v = w + w'$, where $w \in W$ and $w' \in W'$). We say that V is the *direct sum* of W and W', and we write $V = W \oplus W'$.

Corollary 6.49. *Every subspace W of a vector space V is a direct summand.*

Proof. Let B be a basis of W. By the theorem, there is a subset B' with $B \cup B'$ a basis of V. It is straightforward to check that $V = W \oplus \langle B' \rangle$, where $\langle B' \rangle$ denotes the subspace spanned by B'. •

The ring of real numbers $\mathbb{R}$ is a vector space over $\mathbb{Q}$; a basis is usually called a ***Hamel basis***, and it is useful in constructing analytic counterexamples. For example, we may use a Hamel basis to prove the existence of a discontinuous function $f : \mathbb{R} \to \mathbb{R}$ that satisfies the functional equation $f(x + y) = f(x) + f(y)$.[8]

Example 6.50.
An ***inner product*** on a vector space V over a field k is a function

$$V \times V \to k,$$

whose values are denoted by (v, w), such that
(i) $(v + v', w) = (v, w) + (v', w)$ for all $v, v', w \in V$;

[8]Here is a sketch of a proof, using infinite cardinal numbers, that such discontinuous functions f exist. As in the finite-dimensional case, if B is a basis of a vector space V, then any function $f : B \to V$ extends to a linear transformation $F : V \to V$ (see Proposition 7.49); namely, $F(\sum r_i b_i) = \sum r_i f(b_i)$. A Hamel basis has cardinal $c = |\mathbb{R}|$, and so there are $c^c = 2^c > c$ functions $f : \mathbb{R} \to \mathbb{R}$ satisfying the functional equation, for every linear transformation is additive. On the other hand, every continuous function on $\mathbb{R}$ is determined by its values on $\mathbb{Q}$, which is countable. It follows that there are only c continuous functions on $\mathbb{R}$. Therefore, there exist discontinuous functions $f : \mathbb{R} \to \mathbb{R}$ satisfying the functional equation $f(x + y) = f(x) + f(y)$.

We have just proved that there exists a discontinuous $f : \mathbb{R} \to \mathbb{R}$ such that $f(x + y) = f(x) + f(y)$ for all $x, y \in \mathbb{R}$; that is, there is some $a \in \mathbb{R}$ with f discontinuous at a. Thus, there is some $\epsilon > 0$ such that, for every $\delta > 0$, there is a $b \in \mathbb{R}$ with $|b - a| < \delta$ and $|f(b) - f(a)| \geq \epsilon$. Let us show that f is discontinuous at every $c \in \mathbb{R}$. The identity $b - a = (b + c - a) - c$ gives $|(b + c - a) - c| < \delta$, and the identity $f(b + c - a) - f(c) = f(b) - f(a)$ gives $|f(b + c - a) - f(c)| \geq \epsilon$.

(ii) $(\alpha v, w) = \alpha(v, w)$ for all $v, w \in V$ and $\alpha \in k$;

(iii) $(v, w) = (w, v)$ for all $v, w \in V$.

We say that the inner product is ***definite*** if $(v, v) \neq 0$ whenever $v \neq 0$.

We are now going to use a Hamel basis to give a definite inner product on $\mathbb{R}$ all of whose values are rational. Regard $\mathbb{R}$ as a vector space over $\mathbb{Q}$, and let Y be a basis. Using 0 coefficients if necessary, for each $v, w \in \mathbb{R}$, there are $y_i \in Y$ and rationals a_i and b_i with $v = \sum a_i y_i$ and $w = \sum b_i y_i$ (the nonzero a_i and nonzero b_i are uniquely determined by v and w, respectively). Define

$$(v, w) = \sum a_i b_i;$$

note that the sum has only finitely many nonzero terms. It is routine to check that we have defined a definite inner product. ◂

There is a notion of dimension for infinite-dimensional vector spaces; of course, dimension will now be a cardinal number. In the following proof, we shall cite and use several facts about cardinals. We denote the cardinal number of a set X by $|X|$.

Fact I. *Let X and Y be sets, and let $f\colon X \to Y$ be a function. If $f^{-1}(y)$ is finite for every $y \in Y$, then $|X| \leq \aleph_0|Y|$; hence, if Y is infinite, then $|X| \leq |Y|$.*

See Kaplansky, *Set Theory and Metric Spaces*; since X is the disjoint union $X = \bigcup_{y \in Y} f^{-1}(y)$, this result follows from Theorem 16 on page 43.

Fact II. *If X is an infinite set and* $\text{Fin}(X)$ *is the family of all its finite subsets, then* $|\text{Fin}(X)| = |X|$.

See Kaplansky, *Set Theory and Metric Spaces*; this result also follows from Theorem 16 on page 43.

Fact III (Schroeder–Bernstein Theorem). *If X and Y are sets with $|X| \leq |Y|$ and $|Y| \leq |X|$, then $|X| = |Y|$.*

See Birkhoff–Mac Lane, *A Survey of Modern Algebra*, page 387.

Theorem 6.51. *Let k be a field and let V be a vector space over k.*

(i) *Any two bases of V have the same number of elements* (*that is, they have the same cardinal number*); *this cardinal is called the **dimension** of V and it is denoted by* $\dim(V)$.

(ii) *Vector spaces V and V' over k are isomorphic if and only if* $\dim(V) = \dim(V')$.

Proof. (i) Let B and B' be bases of V. If B is finite, then V is finite-dimensional, and hence B' is also finite (Corollary 3.90); moreover, we have proved, in Theorem 3.85, that $|B| = |B'|$. Therefore, we may assume that both B and B' are infinite.

Each $v \in V$ has a unique expression of the form $v = \sum_{b \in B} \alpha_b b$, where $\alpha_b \in k$ and almost all $\alpha_b = 0$. Define the ***support*** of v (with respect to B) by

$$\text{supp}(v) = \{b \in B : \alpha_b \neq 0\};$$

thus, $\text{supp}(v)$ is a finite subset of B for every $v \in V$. Define $f: B' \to \text{Fin}(B)$ by $f(b') = \text{supp}(b')$. Note that if $\text{supp}(b') = \{b_1, \ldots, b_n\}$, then $b' \in \langle b_1, \ldots, b_n \rangle = \langle \text{supp}(b') \rangle$, the subspace spanned by $\text{supp}(b')$. Since $\langle \text{supp}(b') \rangle$ has dimension n, it contains at most n elements of B', because B' is independent (Corollary 3.88). Therefore, $f^{-1}(T)$ is finite for every finite subset T of B [of course, $f^{-1}(T) = \varnothing$ is possible]. By Fact I, we have $|B'| \leq |\text{Fin}(B)|$, and by Fact II, we have $|B'| \leq |B|$. Interchanging the roles of B and B' gives the reverse inequality $|B| \leq |B'|$, and so Fact III gives $|B| = |B'|$.

(ii) Adapt the proof of Corollary 3.105, the finite-dimensional version. •

The next application is a characterization of noetherian rings in terms of their prime ideals.

Lemma 6.52. *Let R be a commutative ring and let $\mathcal{F}$ be the family of all those ideals in R that are not finitely generated. If $\mathcal{F} \neq \varnothing$, then $\mathcal{F}$ has a maximal element.*

Proof. Partially order $\mathcal{F}$ by inclusion. It suffices, by Zorn's lemma, to prove that if $\mathcal{C}$ is a chain in $\mathcal{F}$, then $I^* = \bigcup_{I \in \mathcal{C}} I$ is not finitely generated. If, on the contrary, $I^* = (a_1, \ldots, a_n)$, then $a_j \in I_j$ for some $I_j \in \mathcal{C}$. But $\mathcal{C}$ is a chain, and so one of the ideals $I_1, \ldots, I_n$, call it I_0, contains the others, by Proposition 6.45. It follows that $I^* = (a_1, \ldots, a_n) \subseteq I_0$. The reverse inclusion is clear, for $I \subseteq I^*$ for all $I \in \mathcal{C}$. Therefore, $I_0 = I^*$ is finitely generated, contradicting $I_0 \in \mathcal{F}$. •

Theorem 6.53 (I. S. Cohen). *A commutative ring R is noetherian if and only if every prime ideal in R is finitely generated.*

Proof. Only sufficiency needs proof. Assume that every prime ideal is finitely generated. Let $\mathcal{F}$ be the family of all ideals in R that are not finitely generated. If $\mathcal{F} \neq \varnothing$, then the lemma provides an ideal I that is not finitely generated and that is maximal such. We will show that I is a prime ideal; with the hypothesis that every prime ideal is finitely generated, this contradiction will show that $\mathcal{F} = \varnothing$; that is, that R is noetherian.

Suppose that $ab \in I$ but $a \notin I$ and $b \notin I$. Since $a \notin I$, the ideal $I + Ra$ is strictly larger than I, and so $I + Ra$ is finitely generated; indeed, we may assume that

$$I + Ra = (i_1 + r_1 a, \ldots, i_n + r_n a),$$

where $i_k \in I$ and $r_k \in R$ for all k. Consider $J = (I : a) = \{x \in R : xa \in I\}$. Now $I + Rb \subseteq J$; since $b \notin I$, we have $I \subsetneq J$, and so J is finitely generated. We claim that $I = (i_1, \ldots, i_n, Ja)$. Clearly, $(i_1, \ldots, i_n, Ja) \subseteq I$, for every $i_k \in I$ and $Ja \subseteq I$. For the reverse inclusion, if $z \in I \subseteq I + Ra$, there are $u_k \in R$ with $z = \sum_k u_k(i_k + r_k a)$. Then $(\sum_k u_k r_k)a = z - \sum_k u_k i_k \in I$, so that $\sum_k u_k r_k \in J$. Hence, $z = \sum_k u_k i_k + (\sum_k u_k r_k)a \in (i_1, \ldots, i_n, Ja)$. It follows that $I = (i_1, \ldots, i_n, Ja)$ is finitely generated, a contradiction, and so I is a prime ideal. •

W. Krull has proved that every noetherian ring has the DCC on prime ideals (see Corollary 11.163).

Our next application involves algebraic closures of fields. Recall that a field extension K/k is *algebraic* if every $a \in K$ is a root of some nonzero polynomial $f(x) \in k[x]$; that is, K/k is an algebraic extension if every element $a \in K$ is algebraic over k.

We have already discussed algebraic extensions in Proposition 3.117 on page 185, and the following proposition will add a bit more.

Proposition 6.54. *Let K/k be an extension.*

(i) *If $z \in K$, then z is algebraic over k if and only if $k(z)/k$ is finite.*

(ii) *If $z_1, z_2, \ldots, z_n \in K$ are algebraic over k, then $k(z_1, z_2, \ldots, z_n)/k$ is a finite extension.*

(iii) *If $y, z \in K$ are algebraic over k, then $y + z$, yz, and y^{-1} (for $y \neq 0$) are also algebraic over k.*

(iv) *Define*

$$K_{\text{alg}} = \{z \in K : z \text{ is algebraic over } k\}.$$

Then K_{alg} is a subfield of K.

Proof. (i) If $k(z)/k$ is finite, then Proposition 3.117(i) shows that z is algebraic over k. Conversely, if z is algebraic over k, then Proposition 3.117(v) shows that $k(z)/k$ is finite.

(ii) We prove this by induction on $n \geq 1$; the base step is part (i). For the inductive step, there is a tower of fields

$$k \subseteq k(z_1) \subseteq k(z_1, z_2) \subseteq \cdots \subseteq k(z_1, \ldots, z_n) \subseteq k(z_1, \ldots, z_{n+1}).$$

Now $[k(z_{n+1}) : k]$ is finite, and we have $[k(z_1, \ldots, z_n) : k]$ finite, by the inductive hypothesis. Indeed, $[k(z_{n+1}) : k] = d$, where d is the degree of the monic irreducible polynomial in $k[x]$ having z_{n+1} as a root (by Proposition 3.117). But if z_{n+1} satisfies a polynomial of degree d over k, then it satisfies a polynomial of degree $d' \leq d$ over the larger field $F = k(z_1, \ldots, z_n)$. We conclude that

$$[k(z_1, \ldots, z_{n+1}) : k(z_1, \ldots, z_n)] = [F(z_{n+1}) : F] \leq [k(z_{n+1}) : k].$$

Therefore,

$$[k(z_1, \ldots, z_{n+1}) : k] = [F(z_{n+1}) : k] = [F(z_{n+1}) : F][F : k]$$

is finite.

(iii) Now $k(y, z)/k$ is finite, by part (ii). Therefore, $k(y + z) \subseteq k(y, z)$ and $k(yz) \subseteq k(y, z)$ are also finite, for any subspace of a finite-dimensional vector space is itself finite-dimensional [Corollary 3.90(i)]. By part (i), $y + z$, yz, and y^{-1} are algebraic over k.

(iv) This follows at once from part (iii). •

Definition. Given the extension $\mathbb{C}/\mathbb{Q}$, define the ***algebraic numbers*** by

$$\mathbb{A} = \mathbb{C}_{\text{alg}}.$$

Thus, $\mathbb{A}$ consists of all those complex numbers that are roots of nonzero polynomials in $\mathbb{Q}[x]$, and the proposition shows that $\mathbb{A}$ is a subfield of $\mathbb{C}$ that is algebraic over $\mathbb{Q}$.

Example 6.55.
We claim that $\mathbb{A}/\mathbb{Q}$ is an algebraic extension that is not finite. Suppose, on the contrary, that $[\mathbb{A} : \mathbb{Q}] = n$, for some integer n. Now there exists an irreducible polynomial $p(x) \in \mathbb{Q}[x]$ of degree $n + 1$; for example, take $p(x) = x^{n+1} - 2$. If α is a root of $p(x)$, then $\alpha \in \mathbb{A}$, and so $\mathbb{Q}(\alpha) \subseteq \mathbb{A}$. Thus, $\mathbb{A}$ is an n-dimensional vector space over $\mathbb{Q}$ containing an $(n+1)$-dimensional subspace, and this is a contradiction. ◂

Lemma 6.56.

(i) *If $k \subseteq K \subseteq E$ is a tower of fields with E/K and K/k algebraic, then E/k is also algebraic.*

(ii) *Let*

$$K_0 \subseteq K_1 \subseteq \cdots \subseteq K_n \subseteq K_{n+1} \subseteq \cdots$$

be an ascending tower of fields; if K_{n+1}/K_n is algebraic for all $n \geq 0$, then $K^* = \bigcup_{n\geq 0} K_n$ is a field that is algebraic over K_0.

(iii) *Let $K = k(A)$; that is, K is obtained from k by adjoining the elements in a set A. If each element $a \in A$ is algebraic over k, then K/k is an algebraic extension.*

Proof. (i) Let $e \in E$; since E/K is algebraic, there is some $f(x) = \sum_{i=0}^{n} a_i x^i \in K[x]$ having e as a root. If $F = k(a_0, \ldots, a_n)$, then e is algebraic over F, and so $k(a_0, \ldots, a_n, e) = F(e)$ is a finite extension of F; that is, $[F(e) : F]$ is finite. Since K/k is an algebraic extension, each a_i is algebraic over k, and Corollary 3.90 on page 170 shows that the intermediate field F is finite-dimensional over k; that is, $[F : k]$ is finite.

$$[k(a_0, \ldots, a_n, e) : k] = [F(e) : k] = [F(e) : F][F : k]$$

is finite, and so e is algebraic over k, by Proposition 6.54(i). We conclude that E/k is algebraic.

(ii) If $y, z \in K^*$, then they are there because $y \in K_m$ and $z \in K_n$; we may assume that $m \leq n$, so that both $y, z \in K_n \subseteq K^*$. Since K_n is a field, it contains $y + z$, yz, and y^{-1} if $y \neq 0$. Therefore, K^* is a field.

If $z \in K^*$, then z must lie in K_n for some n. But K_n/K_0 is algebraic, by an obvious inductive generalization of part (i), and so z is algebraic over K_0. Since every element of K^* is algebraic over K_0, the extension K^*/K_0 is algebraic.

(iii) Let $z \in k(A)$; by Exercise 3.95 on page 197, there is an expression for z involving k and finitely many elements of A; say, $a_1, \ldots, a_m$. Hence, $z \in k(a_1, \ldots, a_m)$. By Proposition 6.54(ii), $k(z)/k$ is finite and hence z is algebraic over k. •

Definition. A field K is ***algebraically closed*** if every nonconstant $f(x) \in K[x]$ has a root in K. An ***algebraic closure*** of a field k is an algebraic extension $\overline{k}$ of k that is algebraically closed.

The algebraic closure of $\mathbb{Q}$ turns out to be the algebraic numbers: $\overline{\mathbb{Q}} = \mathbb{A}$. The fundamental theorem of algebra says that $\mathbb{C}$ is algebraically closed; moreover, $\mathbb{C}$ is an algebraic closure of $\mathbb{R}$. We have already given an algebraic proof, Theorem 4.49, but perhaps the simplest proof of this theorem is by Liouville's theorem in complex variables: Every bounded entire function is constant. If $f(x) \in \mathbb{C}[x]$ had no roots, then $1/f(x)$ would be a bounded entire function that is not constant.

There are two main results here. First, every field has an algebraic closure; second, any two algebraic closures of a field are isomorphic. Our proof of existence will make use of a "big" polynomial ring: We assume that if k is a field and T is an infinite set, then there is a polynomial ring $k[T]$ having one variable for each $t \in T$. (We have already constructed $k[T]$ when T is finite, and the infinite case is essentially a union of $k[U]$, where U ranges over all the finite subsets of T. A construction of $k[T]$ for infinite T will be given in Exercise 9.93 on page 756.)

Lemma 6.57. *Let k be a field, and let $k[T]$ be the polynomial ring in a set T of variables. If $t_1, \ldots, t_n \in T$ are distinct and if $f_i(t_i) \in k[t_i] \subseteq k[T]$ are nonconstant polynomials, then the ideal $I = (f_1(t_1), \ldots, f_n(t_n))$ in $k[T]$ is a proper ideal.*

Remark. If $n = 2$, then $f_1(t_1)$ and $f_2(t_2)$ are relatively prime, and this lemma says that 1 is not a linear combination of them. ◂

Proof. If I is not a proper ideal in $k[T]$, then there exist $h_i(T) \in k[T]$ with

$$1 = h_1(T) f_1(t_1) + \cdots + h_n(T) f_n(t_n).$$

Consider the field extension $k(\alpha_1, \ldots, \alpha_n)$, where α_i is a root of $f_i(t_i)$ for $i = 1, \ldots, n$ (the f_i are not constant). Denote the variables involved in the $h_i(T)$ other than $t_1, \ldots, t_n$, if any, by $t_{n+1}, \ldots, t_m$. Evaluating when $t_i = \alpha_i$ if $i \leq n$ and $t_i = 0$ if $i \geq n+1$ (evaluation is a ring homomorphism $k[T] \to k(\alpha_1, \ldots, \alpha_n)$), the right side is 0, and we have the contradiction $1 = 0$. •

Theorem 6.58. *Given a field k, there exists an algebraic closure $\overline{k}$ of k.*

Proof. Let T be a set in bijective correspondence with the family of nonconstant polynomials in $k[x]$. Let $R = k[T]$ be the big polynomial ring, and let I be the ideal in R generated by all elements of the form $f(t_f)$, where $t_f \in T$; that is, if

$$f(x) = x^n + a_{n-1}x^{n-1} + \cdots + a_0,$$

where $a_i \in k$, then

$$f(t_f) = (t_f)^n + a_{n-1}(t_f)^{n-1} + \cdots + a_0.$$

We claim that the ideal I is proper; if not, $1 \in I$, and there are distinct $t_1, \ldots, t_n \in T$ and polynomials $h_1(T), \ldots, h_n(T) \in k[T]$ with $1 = h_1(T) f_1(t_1) + \cdots + h_n(T) f_n(t_n)$, contradicting the lemma. Therefore, there is a maximal ideal M in R containing I, by Theorem 6.46. Define $K = R/M$. The proof is now completed in a series of steps.

(i) *K/k is a field extension.*

We know that $K = R/M$ is a field because M is a maximal ideal. Moreover, the ring map θ, which is the composite

$$k \xrightarrow{i} k[T] = R \xrightarrow{\text{nat}} R/M = K,$$

(where i is the inclusion) is not identically 0 because $1 \mapsto 1$, and hence θ is injective, by Corollary 3.53. We identify k with $\operatorname{im}\theta \subseteq K$.

(ii) *Every nonconstant $f(x) \in k[x]$ splits in $K[x]$.*

By definition, there is $t_f \in T$ with $f(t_f) \in I \subseteq M$, and the coset $t_f + M \in R/M = K$ is a root of $f(x)$. It now follows by induction on degree that $f(x)$ splits over K.

(iii) *The extension K/k is algebraic.*

By Lemma 6.56(iii), it suffices to show that each $t_f + M$ is algebraic over k [for $K = k(\text{all } t_f + M)$]; but this is obvious, for t_f is a root of $f(x) \in k[x]$.

(iv) *K is algebraically closed*

Let $g(x) \in K[x]$ and let $E = K(\alpha_1, \ldots, \alpha_m)$ be a splitting field of $g(x)$ over K. We have a tower of fields $k \subseteq K \subseteq E$ in which K/k and E/K are algebraic extensions. By Lemma 6.56(i), E/k is an algebraic extension. Hence, $p(x) = \operatorname{irr}(\alpha_1, k) \in k[x]$. By item (ii), $p(x)$ splits over K, so that $\{\alpha_1, \ldots, \alpha_m\} \subseteq K$; that is, $E \subset K$. Therefore, $g(x)$ splits in $K[x]$, and so K is algebraically closed. •

Corollary 6.59. *If k is a countable field, then it has a countable algebraic closure. In particular, an algebraic closure of $\mathbb{Q}$ or of $\mathbb{F}_p$ is countable.*

Proof. If k is countable, then the set T of all nonconstant polynomials is countable, say, $T = \{t_1, t_2, \ldots\}$, because $k[x]$ is countable. Hence, $k[T] = \bigcup_{\ell \geq 1} k[t_1, \ldots, t_\ell]$ is countable, as is its quotient k_1 (in the proof of Theorem 6.58). It follows, by induction on $n \geq 0$, that every k_n is countable. Finally, a countable union of countable sets is itself countable, so that an algebraic closure of k is countable. •

We are now going to prove the uniqueness of an algebraic closure.

Definition. If F/k and K/k are extensions, then a ***k-map*** is a ring homomorphism $\varphi : F \to K$ that fixes k pointwise.

We note that if K/k is an extension, if $\varphi : K \to K$ is a k-map, and if $a \in K$ is a root of some irreducible polynomial $p(x) \in k[x]$, then φ permutes all the roots $\{a = a_1, a_2, \ldots, a_r\}$ of $p(x)$ that lie in K. If $p(x) = x^n + c_{n-1}x^{n-1} + \cdots + c_0$, then

$$0 = p(a_i) = a_i^n + c_{n-1}a_i^{n-1} + \cdots + c_0,$$

and so

$$\begin{aligned} 0 &= [\varphi(a_i)]^n + \varphi(c_{n-1})[\varphi(a_i)]^{n-1} + \cdots + \varphi(c_0) \\ &= [\varphi(a_i)]^n + c_{n-1}[\varphi(a_i)]^{n-1} + \cdots + c_0, \end{aligned}$$

because φ fixes all $c_i \in k$. Therefore, $\varphi(a_i)$ is a root of $p(x)$ lying in K. Finally, since φ is injective and $\{a_1, \ldots, a_r\}$ is finite, φ is a permutation of these roots.

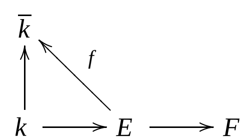
Lemma 6.60. *If K/k is an algebraic extension, then every k-map $\varphi\colon K \to K$ is an automorphism of K.*

Proof. By Corollary 3.53, the k-map φ is injective. To see that φ is surjective, let $a \in K$. Since K/k is algebraic, there is an irreducible polynomial $p(x) \in k[x]$ having a as a root. As we remarked earlier, φ being a k-map implies that it permutes the set of those roots of $p(x)$ that lie in K. Therefore, $a \in \operatorname{im}\varphi$ because $a = \varphi(a_i)$ for some i. •

The next lemma will use Zorn's lemma by partially ordering a family of functions. Since a function is essentially a set, its graph, it is reasonable to take a union of functions in order to obtain an upper bound; we give details below.

Lemma 6.61. *If $\overline{k}/k$ is an algebraic closure, and if F/k is an algebraic extension, then there is an injective k-map $\psi : F \to \overline{k}$.*

Proof. If E is an intermediate field, $k \subseteq E \subseteq F$, let us call an ordered pair (E, f) an "approximation" if $f : E \to \overline{k}$ is a k-map. In the following diagram, all arrows other than f are inclusions.



Define

$$X = \{\text{approximations } (E, f)\}.$$

Note that $X \neq \varnothing$ because $(k, 1_k) \in X$. Partially order X by

$$(E, f) \preceq (E', f') \quad \text{if} \quad E \subseteq E' \text{ and } f'|E = f.$$

That the restriction $f'|E$ is f means that f' *extends* f; that is, both functions agree whenever possible: $f'(u) = f(u)$ for all $u \in E$.

It is easy to see that an upper bound of a chain

$$\mathcal{S} = \{(E_j, f_j) : j \in J\}$$

is given by $(\bigcup E_j, \bigcup f_j)$. That $\bigcup E_j$ is an intermediate field is, by now, a routine argument. We can take the union of the graphs of the f_j's, but here is a more down-to-earth description of $\Phi = \bigcup f_j$: If $u \in \bigcup E_j$, then $u \in E_{j_0}$ for some j_0, and $\Phi\colon u \mapsto f_{j_0}(u)$.

Note that Φ is well-defined: If $u \in E_{j_1}$, we may assume, for notation, that $E_{j_0} \subseteq E_{j_1}$, and then $f_{j_1}(u) = f_{j_0}(u)$ because f_{j_1} extends f_{j_0}. The reader may check that Φ is a k-map.

By Zorn's lemma, there exists a maximal element (E_0, f_0) in X. We claim that $E_0 = F$, and this will complete the proof (take $\psi = f_0$). If $E_0 \subsetneq F$, then there is $a \in F$ with $a \notin E_0$. Since F/k is algebraic, we have F/E_0 algebraic, and there is an irreducible $p(x) \in E_0[x]$ having a as a root; since $\overline{k}/k$ is algebraic and $\overline{k}$ is algebraically closed, we have a factorization in $\overline{k}[x]$:

$$f_0^*(p(x)) = \prod_{i=1}^{n}(x - b_i),$$

where $f_0^*\colon E_0[x] \to \overline{k}[x]$ is the map induced by f_0. If all the b_i lie in $f_0(E_0) \subseteq \overline{k}$, then $f_0^{-1}(b_i) \in E_0 \subseteq F$ for all i, and there is a factorization of $p(x)$ in $F[x]$, namely, $p(x) = \prod_{i=1}^{n}[x - f_0^{-1}(b_i)]$. But $a \notin E_0$ implies $a \neq f_0^{-1}(b_i)$ for any i. Thus, $x - a$ is another factor of $p(x)$ in $F[x]$, contrary to unique factorization. We conclude that there is some $b_i \notin \operatorname{im} f_0$. By Theorem 3.120(ii), we may define $f_1\colon E_0(a) \to \overline{k}$ by

$$c_0 + c_1 a + c_2 a^2 + \cdots \mapsto f_0(c_0) + f_0(c_1) b_i + f_0(c_2) b_i^2 + \cdots .$$

A straightforward check shows that f_1 is a (well-defined) k-map extending f_0. Hence, $(E_0, f_0) \prec (E_0(a), f_1)$, contradicting the maximality of (E_0, f_0). This completes the proof. •

Theorem 6.62. *Any two algebraic closures of a field k are isomorphic via a k-map.*

Proof. Let K and L be two algebraic closures of a field k. By Lemma 6.61, there are k-maps $\psi\colon K \to L$ and $\theta\colon L \to K$. By Lemma 6.60, both composites $\theta\psi\colon K \to K$ and $\psi\theta\colon L \to L$ are automorphisms. It follows that ψ (and θ) is a k-isomorphism. •

It is now permissible to speak of *the* algebraic closure of a field.

In the remainder of this section, we investigate the structure of arbitrary fields; we begin with ***simple transcendental extensions*** $k(x)$, where k is a field and x is transcendental over k; that is, we examine the function field $k(x)$.

Definition. If $\varphi \in k(x)$, then there are polynomials $g(x), h(x) \in k[x]$ with $(g, h) = 1$ and $\varphi = g(x)/h(x)$. Define the ***degree*** of φ by

$$\text{degree}(\varphi) = \max\{\deg(g), \deg(h)\}.$$

A rational function $\varphi \in k(x)$ is called a ***linear fractional transformation*** if

$$\varphi = \frac{ax + b}{cx + d},$$

where $a, b, c, d \in k$ and $ad - bc \neq 0$.

Now $\varphi \in k(x)$ has degree 0 if and only if φ is a constant (that is, $\varphi \in k$), while Exercise 6.56 on page 375 says that $\varphi \in k(x)$ has degree 1 if and only if φ is a linear fractional transformation. If $A = \left[\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right] \in \mathrm{GL}(2, k)$, write $\langle A \rangle = (ax + b)/(cx + d)$. If we define $\langle A' \rangle \langle A \rangle = \langle A'A \rangle$, then it is easily checked that the set $\mathrm{LF}(k)$ of all linear fractional transformations with entries in k is a group under this operation. In Exercise 6.57 on page 375, the reader will prove that $\mathrm{LF}(k) \cong \mathrm{PGL}(2, k) = \mathrm{GL}(2, k)/Z(2, k)$, where $Z(2, k)$ is the (normal) subgroup of all 2×2 (nonzero) scalar matrices.

Proposition 6.63. *If $\varphi \in k(x)$ is nonconstant, then φ is transcendental over k and $k(x)/k(\varphi)$ is a finite extension with*

$$[k(x) : k(\varphi)] = \mathrm{degree}(\varphi).$$

Moreover, if $\varphi = g(x)/h(x)$ and $(g, h) = 1$, then

$$\mathrm{irr}(x, k(\varphi)) = g(y) - \varphi h(y).$$

Proof. Let $g(x) = \sum a_i x^i$ and $h(x) = \sum b_i x^i \in k[x]$. Now $\theta(y) = g(y) - \varphi h(y)$ is a polynomial in $k(\varphi)[y]$:

$$\theta(y) = \sum a_i y^i - \varphi \sum b_i y^i = \sum (a_i - \varphi b_i) y^i.$$

If $\theta(y)$ were the zero polynomial, then all its coefficients would be 0. But if b_i is a nonzero coefficient of $h(y)$, then $a_i - \varphi b_i = 0$ gives $\varphi = a_i/b_i$, contradicting the assumption that φ is not a constant; that is, $\varphi \notin k$. It follows that

$$\deg(\theta) = \deg(g(y) - \varphi h(y)) = \max\{\deg(g), \deg(h)\} = \mathrm{degree}(\varphi).$$

Since x is a root of $\theta(y)$, we have x algebraic over $k(\varphi)$. If φ were algebraic over k, then $k(\varphi)/k$ would be finite, giving $[k(x) : k] = [k(x) : k(\varphi)][k(\varphi) : k]$ finite, a contradiction. Therefore, φ is transcendental over k.

We claim that $\theta(y)$ is an irreducible polynomial in $k(\varphi)[y]$. If not, then $\theta(y)$ factors in $k[\varphi][y]$, by Gauss's Corollary 6.27. But $\theta(y) = g(y) - \varphi h(y)$ is linear in φ, and so Corollary 6.37 shows that $\theta(y)$ is irreducible. Finally, since $\deg(\theta) = \mathrm{degree}(\varphi)$, we have $[k(x) : k(\varphi)] = \mathrm{degree}(\varphi)$. •

Corollary 6.64. *Let $\varphi \in k(x)$, where $k(x)$ is the field of rational functions over a field k. Then $k(\varphi) = k(x)$ if and only if φ is a linear fractional transformation.*

Proof. By Proposition 6.63, $k(\varphi) = k(x)$ if and only if $\mathrm{degree}(\varphi) = 1$; that is, φ is a linear fractional transformation. •

Corollary 6.65. *If $k(x)$ is the field of rational functions over a field k, then*

$$\mathrm{Gal}(k(x)/k) \cong \mathrm{LF}(k),$$

the group of all linear fractional transformations over k.

Proof. Let $\sigma\colon k(x) \to k(x)$ be an automorphism of $k(x)$ fixing k. Now $\sigma\colon x \mapsto x^\sigma$, where $x^\sigma \in k(x)$; since σ is surjective, we must have $k(x^\sigma) = k(x)$, and so x^σ is a linear fractional transformation, by Corollary 6.64. Define $\gamma\colon \mathrm{Gal}(k(x)/k) \to \mathrm{LF}(k)$ by $\gamma\colon \sigma \mapsto x^\sigma$. The reader may check that γ is a homomorphism ($x^{\sigma\tau} = x^\tau x^\sigma$); γ is an isomorphism because γ^{-1} is the function assigning, to any linear fractional transformation $\varphi = (ax+b)/(cx+d)$, the automorphism of $k(x)$ that sends x to φ. •

Theorem 6.66 (Lüroth's Theorem). *If $k(x)$ is a simple transcendental extension, then every intermediate field B is also a simple transcendental extension of k: There is $\varphi \in B$ with $B = k(\varphi)$.*

Proof. If $\beta \in B$ is not constant, then $[k(x) : k(\beta)] = [k(x) : B][B : k(\beta)]$ is finite, by Proposition 6.63; hence, $[k(x) : B]$ is finite and x is algebraic over B. The proof of Proposition 6.63 shows that if $\varphi \in k(x)$, then φ is a coefficient of $\mathrm{irr}(x, k(\varphi))$; the proof of Lüroth's theorem is a converse, showing that $B = k(\varphi)$ for some coefficient φ of $\mathrm{irr}(x, B)$. Now

$$\mathrm{irr}(x, B) = y^n + \beta_{n-1}y^{n-1} + \cdots + \beta_0 \in B[y].$$

Each coefficient $\beta_\ell \in B \subseteq k(x)$ is a rational function, which we write in lowest terms: $\beta_\ell = g_\ell(x)/h_\ell(x)$, where $g_\ell(x), h_\ell(x) \in k[x]$ and $(g_\ell, h_\ell) = 1$. As in Lemma 6.24(i), the content $c(\mathrm{irr}) = d(x)/b(x)$, where $b(x)$ is the product of the h_ℓ and $d(x)$ is their gcd. It is easy to see that $f(x)$, defined by $f(x) = b(x)/d(x)$, lies in $k[x]$; in fact, the reader may generalize Exercise 1.26 on page 13 to show that $f(x)$ is the lcm of the h_ℓ. Define

$$i(x, y) = f(x)\,\mathrm{irr}(x, B),$$

the associated primitive polynomial in $k[x][y]$ (of course, $k[x][y] = k[x, y]$, but we wish to view it as polynomials in y with coefficients in $k[x]$). If we denote the highest exponent of y occurring in a polynomial $a(x, y)$ by $\deg_y(a)$, then $n = \deg_y(i)$; let $m = \deg_x(i)$. Since $i(x, y) = f(x)y^n + \sum_{\ell=0}^{n-1} f(x)\beta_\ell y^\ell$, we have $m = \max_\ell\{\deg(f), \deg(f\beta_\ell)\}$. Now $h_\ell(x) \mid f(x)$ for all ℓ, so that $\deg(h_\ell) \le \deg(f) \le m$ [because $f(x)$ is one of the coefficients of $i(x, y)$]. Also,

$$f\beta_\ell = \frac{h_0 \cdots h_{n-1}}{d} \cdot \frac{g_\ell}{h_\ell} = \frac{h_0 \cdots \widehat{h_\ell} \cdots h_{n-1}}{d} g_\ell.$$

Since $(h_0 \cdots \widehat{h_\ell} \cdots h_{n-1})/d \in k[x]$, we have $\deg(g_\ell) \le \deg(f\beta_\ell) \le m$. We conclude that $\deg(g_\ell) \le m$ and $\deg(h_\ell) \le m$.

Some coefficient β_j of $\operatorname{irr}(x, B)$ is not constant, lest x be algebraic over k. Omit the subscripts j, write $\beta_j = g(x)/h(x)$, and define

$$\varphi = \beta_j = g(x)/h(x) \in B.$$

Now $g(y) - \varphi h(y) = g(y) - g(x)h(x)^{-1}h(y) \in B[y]$ has x as a root, and so $\operatorname{irr}(x, B)$ divides $g(y) - \varphi h(y)$ in $B[y] \subseteq k(x)[y]$. Therefore, there is $q(x, y) \in k(x)[y]$ with

$$\operatorname{irr}(x, B)q(x, y) = g(y) - \varphi h(y). \tag{1}$$

Since $g(y) - \varphi h(y) = h(x)^{-1}\big(h(x)g(y) - g(x)h(y)\big)$, the content $c\big(g(y) - \varphi h(y)\big)$ is $h(x)^{-1}$ and the associated primitive polynomial is

$$\Phi(x, y) = h(x)g(y) - g(x)h(y).$$

Notice that $\Phi(x, y) \in k[x][y]$ and that $\Phi(y, x) = -\Phi(x, y)$.

Rewrite Eq. (1), where $c(q) \in k(x)$ is the content of $q(x, y)$:

$$f(x)^{-1}i(x, y)c(q)q(x, y)^*h(x) = \Phi(x, y)$$

(remember that $f(x)^{-1}$ is the content of $\operatorname{irr}(x, B)$ and $i(x, y)$ is its associated primitive polynomial). The product $i(x, y)q(x, y)^*$ is primitive, by Gauss's Lemma 6.23. But $\Phi(x, y) \in k[x][y]$, so that Lemma 6.24(iii) gives $f(x)^{-1}c(q)h(x) \in k[x]$. We now define $q^{**}(x, y) = f(x)^{-1}c(q)h(x)q(x, y)$, so that $q^{**}(x, y) \in k[x, y]$ and

$$i(x, y)q^{**}(x, y) = \Phi(x, y) \quad \text{in } k[x, y]. \tag{2}$$

Let us compute degrees in Eq. (2): the degree in x of the left hand side is

$$\deg_x(iq^{**}) = \deg_x(i) + \deg_x(q^{**}) = m + \deg_x(q^{**}), \tag{3}$$

while the degree in x of the right hand side is

$$\deg_x(\Phi) = \max\{\deg(g), \deg(h)\} \le m, \tag{4}$$

as we saw above. We conclude that $m + \deg_x(q^{**}) \le m$, so that $\deg_x(q^{**}) = 0$; that is, $q^{**}(x, y)$ is a function of y alone. But $\Phi(x, y)$ is a primitive polynomial in x, and hence the symmetry $\Phi(y, x) = -\Phi(x, y)$ shows that it is also a primitive polynomial in y. Thus, q^{**} is a constant, and so $i(x, y)$ and $\Phi(x, y)$ are associates in $k[x, y]$; hence, $\deg_x(\Phi) = \deg_x(i) = m$. With Eq. (4), this equality gives

$$m = \deg_x(\Phi) = \max\{\deg(g), \deg(h)\}.$$

Symmetry of Φ also gives $\deg_y(\Phi) = \deg_x(\Phi)$, and so

$$n = \deg_y(\Phi) = \deg_x(\Phi) = m = \max\{\deg(g), \deg(h)\}.$$

By definition, $\operatorname{degree}(\varphi) = \max\{\deg(g), \deg(h)\} = m$; hence, Proposition 6.63 gives $[k(x) : k(\varphi)] = m$. Finally, since $\varphi \in B$, we have $[k(x) : k(\varphi)] = [k(x) : B][B : k(\varphi)]$. As $[k(x) : B] = n = m$, this forces $[B : k(\varphi)] = 1$; that is, $B = k(\varphi)$. •

There are examples of intermediate fields B with $k \subseteq B \subseteq k(x_1, \ldots, x_n)$, for $n > 1$, that are not so easily described.

We now consider more general field extensions.

Definition. Let E/k be a field extension. A subset U of E is ***algebraically dependent*** over k if there exists a finite subset $\{u_1, \ldots, u_n\} \subseteq U$ and a nonzero polynomial $f(x_1, \ldots, x_n) \in k[x_1, \ldots, x_n]$ with $f(u_1, \ldots, u_n) = 0$. A subset B of E is ***algebraically independent*** if it is not algebraically dependent.

Let E/k be a field extension, let $u_1, \ldots, u_n \in E$, and let $\varphi\colon k[x_1, \ldots, x_n] \to E$ be the evaluation map; that is, φ is the homomorphism sending $f(x_1, \ldots, x_n)$ to $f(u_1, \ldots, u_n)$ for all $f(x_1, \ldots, x_n) \in k[x_1, \ldots, x_n]$. Now $\{u_1, \ldots, u_n\}$ is algebraically dependent if and only if $\ker \varphi \neq \{0\}$. If $\{u_1, \ldots, u_n\}$ is algebraically independent, then φ extends to an isomorphism $k(x_1, \ldots, x_n) \cong k(u_1, \ldots, u_n) \subseteq E$, where $k(x_1, \ldots, x_n)$ is the field of rational functions $\mathrm{Frac}(k[x_1, \ldots, x_n])$. In particular, $\{x_1, \ldots, x_n\} \subseteq E = k(x_1, \ldots, x_n)$ is algebraically independent, for φ is the identity map in this case.

Since algebraically dependent subsets are necessarily nonempty, it follows that the empty subset $\varnothing$ is algebraically independent. A singleton $\{e\} \subseteq E$ is algebraically dependent if e is algebraic over k; that is, e is a root of a nonconstant polynomial over k, and it is algebraically independent if e is transcendental over k, in which case $k(e) \cong k(x)$.

Proposition 6.67. *Let E/k be a field extension. and let $U \subseteq E$. Then U is algebraically dependent over k if and only if there is $u \in U$ with u algebraic over $k(U - \{u\})$.*

Proof. If U is algebraically dependent over k, then there is a finite algebraically dependent subset $U' = \{u_1, \ldots, u_n\} \subseteq U$. We prove, by induction on $n \geq 1$, that some u_i is algebraic over $k(U' - \{u_i\})$. If $n = 1$, then there is some nonzero $f(x) \in k[x]$ with $f(u_1) = 0$; that is, u_1 is algebraic over k. But $U' - \{u_1\} = \varnothing$, and so u_1 is algebraic over $k(U' - \{u_1\}) = k(\varnothing) = k$. For the inductive step, let $U' = \{u_1, \ldots, u_{n+1}\}$ be algebraically dependent. We may assume that $\{u_1, \ldots, u_n\}$ is algebraically independent; otherwise, the inductive hypothesis gives some u_j, for $1 \leq j \leq n$, which is algebraic over $k(u_1, \ldots, \widehat{u}_j, \ldots, u_n)$, and hence, algebraic over $k(U' - \{u_j\})$. Since U' is algebraically dependent, there is a nonzero $f(X, y) \in k[x_1, \ldots, x_n, y]$ with $f(\vec{u}, u_{n+1}) = 0$, where $X = (x_1, \ldots, x_n)$, y is a new variable, and $\vec{u} = (u_1, \ldots, u_n)$. We may write $f(X, y) = \sum_i g_i(X) y^i$, where $g_i(X) \in k[X]$ (because $k[X, y] = k[X][y]$). Since $f(X, y) \neq 0$, some $g_i(X) \neq 0$, and it follows from the algebraic independence of $\{u_1, \ldots, u_n\}$ that $g_i(\vec{u}) \neq 0$. Therefore, $h(y) = \sum_i g_i(\vec{u}) y^i \in k(U)[y]$ is not the zero polynomial. But $0 = f(\vec{u}, u_{n+1}) = h(u_{n+1})$, so that u_{n+1} is algebraic over $k(u_1, \ldots, u_n)$.

For the converse, assume that u is algebraic over $k(U - \{u\})$. We may assume that $U - \{u\}$ is finite, say, $U - \{u\} = \{u_1, \ldots, u_n\}$, where $n \geq 0$ (if $n = 0$, we mean that $U - \{u\} = \varnothing$). We prove, by induction on $n \geq 0$, that U is algebraically dependent. If $n = 0$, then u is algebraic over k, and so $\{u\}$ is algebraically dependent. For the inductive step, let $U - \{u_{n+1}\} = \{u_1, \ldots, u_n\}$. We may assume that $U - \{u_{n+1}\} = \{u_1, \ldots, u_n\}$ is algebraically independent, for otherwise $U - \{u_{n+1}\}$, and hence its superset U, is algebraically dependent. By hypothesis, there is a nonzero polynomial $f(y) = \sum_i c_i y^i \in k(u_1, \ldots, u_n)[y]$ with $f(u_{n+1}) = 0$. As $f(y) \neq 0$, we may assume that one of its terms, say, $c_j \neq 0$. Now $c_i \in k(u_1, \ldots, u_n)$ for each i, and so there are rational functions $c_i(x_1, \ldots, x_n)$ with $c_i(\vec{u}) = c_i$, where $\vec{u} = (u_1, \ldots, u_n)$. Since $f(u_{n+1}) = 0$, we may clear denominators and assume that each $c_i(x_1, \ldots, x_n)$ is a polynomial in $k[x_1, \ldots, x_n]$.

Moreover, $c_j(\vec{u}) \neq 0$ implies $c_j(x_1, \ldots, x_n) \neq 0$, and so

$$g(x_1, \ldots, y) = \sum_i c_i(x_1, \ldots, x_n)y^i$$

is nonzero. Therefore, $\{u_1, \ldots, u_{n+1}\}$ is algebraically dependent. •

Definition. A field extension E/k is ***purely transcendental*** if either $E = k$ or E contains an algebraically independent subset B and $E = k(B)$.

If $X = \{x_1, \ldots, x_n\}$ is a finite set, then

$$k(X) = k(x_1, \ldots, x_n) = \operatorname{Frac}\big(k[x_1, \ldots, x_n]\big)$$

is called the ***function field in*** n ***variables***.

We are going to prove that if E/k is a field extension, then there exists an intermediate field F with F/k purely transcendental and E/F algebraic. In fact, $F = k(B)$, where B is a maximal algebraically independent subset of E/k, and any two such subsets have the same cardinal. The proof is essentially the same as a proof of the invariance of dimension of a vector space, and so we axiomatize that proof.

Recall that a *relation* R from a set Y to a set Z is a subset $R \subseteq Y \times Z$: we write $y \; R \; z$ instead of $(y, z) \in R$. In particular, if Ω is a set, $\mathcal{P}(\Omega)$ is the family of all its subsets, and $\preceq$ is a relation from Ω to $\mathcal{P}(\Omega)$, then we write

$$x \preceq S$$

instead of $(x, S) \in \preceq$.

Definition. A ***dependency relation*** on a set Ω is a relation $\preceq$ from Ω to $\mathcal{P}(\Omega)$ that satisfies the following axioms:

(i) if $x \in S$, then $x \preceq S$;

(ii) if $x \preceq S$, then there exists a finite subset $S' \subseteq S$ with $x \preceq S'$;

(iii) (***Transitivity***) if $x \preceq S$ and if, for some $T \subseteq \Omega$, we have $s \preceq T$ for every $s \in S$, then $x \preceq T$;

(iv) (***Exchange Axiom***) if $x \preceq S$ and $x \not\preceq S - \{y\}$, then $y \preceq (S - \{y\}) \cup \{x\}$.

The transitivity axiom says that if x is dependent on a set S, and if each element of S is dependent on another set T, then x is dependent on T.

Example 6.68.
If Ω is a vector space, then define $x \preceq S$ to mean $x \in \langle S \rangle$, the subspace spanned by S. We claim that $\preceq$ is a dependency relation. The first three axioms are easily checked. We verify the exchange axiom. If $x \preceq S$ and $x \not\preceq S - \{y\}$, then $S = S' \cup \{y\}$ with $y \notin S'$. There are scalars a_i, a with $x = ay + \sum_i a_i s_i$, where $s_i \in S'$; since $x \notin \langle S' \rangle$, we must have $a \neq 0$. Therefore, $y = a^{-1}(x - \sum_i a_i s_i) \in \langle S', x \rangle$, and so $y \preceq S' \cup \{x\}$. ◄

Lemma 6.69. *If E/k is a field extension, then $\alpha \preceq S$, defined by α is algebraic over $k(S)$, is a dependency relation.*

Proof. It is easy to check the first two axioms in the definition of dependency relation, and we now verify axiom (iii): If $x \preceq S$ and if, for some $T \subseteq \Omega$, we have $s \preceq T$ for every $s \in S$, then $x \preceq T$. If F is an intermediate field, denote the field of all $e \in E$ that are algebraic over F by $\overline{F}$. Using this notation, $x \preceq S$ if and only if $x \in \overline{k(S)}$. Moreover, $s \preceq T$ for every $s \in S$ says that $S \subseteq \overline{k(T)}$. It follows that $x \in \overline{k(T)}$, by Lemma 6.56(i), and so $x \preceq T$.

The exchange axiom says, If $u \preceq S$ and $u \not\preceq S - \{v\}$, then $v \preceq (S - \{v\}) \cup \{u\}$. Write $S' = S - \{v\}$, so that u is algebraic over $k(S)$ and u is transcendental over $k(S')$. Now $\{u, v\}$ is algebraically dependent over $k(S')$, by Proposition 6.67, and so there is a nonzero polynomial $f(x, y) \in k(S')[x, y]$ with $f(u, v) = 0$. In more detail, $f(x, y) = g_0(x) + g_1(x)y + \cdots + g_n(x)y^n$, where $g_n(x)$ is nonzero. Since u is transcendental over $k(S')$, we must have $g_n(u) \neq 0$. Therefore, $h(y) = f(u, y) \in k(S', u)[y]$ is a nonzero polynomial. But $h(v) = f(u, v) = 0$, and so v is algebraic over $k(S', u)$; that is, $v \preceq S' \cup \{u\} = (S - \{v\}) \cup \{u\}$. •

Example 6.68 suggests the following terminology.

Definition. Let $\preceq$ be a dependency relation on a set Ω. Call a subset $S \subseteq \Omega$ ***dependent*** if there exists $s \in S$ with $s \preceq S - \{s\}$; call S ***independent*** if it is not dependent. We say that a subset S ***generates*** Ω if $x \preceq S$ for all $x \in \Omega$. A ***basis*** of Ω is an independent subset that generates Ω.

Note that $\varnothing$ is independent, for dependent subsets have elements. If $S \neq \varnothing$, then S is independent if and only if $s \not\preceq S - \{s\}$ for all $s \in S$. It follows that every subset of an independent set is itself independent. By Proposition 6.67, algebraic independence defined on page 361 coincides with independence just defined for the dependency relation in Lemma 6.69.

Lemma 6.70. *Let $\preceq$ be a dependency relation on a set Ω. If $T \subseteq \Omega$ is independent and $z \not\preceq T$ for some $z \in \Omega$, then $T \cup \{z\} \supsetneq T$ is a strictly larger independent subset.*

Proof. Since $z \not\preceq T$, axiom (i) gives $z \notin T$, and so $T \subsetneq T \cup \{z\}$; it follows that $(T \cup \{z\}) - \{z\} = T$. If $T \cup \{z\}$ is dependent, then there exists $t \in T \cup \{z\}$ with $t \preceq (T \cup \{z\}) - \{t\}$. If $t = z$, then $z \preceq T \cup \{z\} - \{z\} = T$, contradicting $z \not\preceq T$. Therefore, $t \in T$. Since T is independent, $t \not\preceq T - \{t\}$. If we set $S = T \cup \{z\} - \{t\}$, $t = x$, and $y = z$ in the exchange axiom, we conclude that $z \preceq (T \cup \{z\} - \{t\}) - \{z\} \cup \{t\} = T$, contradicting the hypothesis $z \not\preceq T$. Therefore, $T \cup \{z\}$ is independent. •

We now generalize the proof of the exchange lemma, Lemma 3.84, and its application to invariance of dimension, Theorem 3.85.

Theorem 6.71. *If $\preceq$ is a dependency relation on a set Ω, then Ω has a basis. In fact, every independent subset B of Ω is part of a basis.*

Proof. Since the empty set $\varnothing$ is independent, the second statement implies the first.

We use Zorn's lemma to prove the existence of maximal independent subsets of Ω containing B. Let X be the family of all independent subsets of Ω containing B, partially ordered by inclusion. Note that X is nonempty, for $B \in X$. Suppose that $\mathcal{C}$ is a chain in X. It is clear that $C^* = \bigcup_{C \in \mathcal{C}}$ is an upper bound of $\mathcal{C}$ if it lies in X; that is, if C^* is independent. If, on the contrary, C^* is dependent, then there is $y \in C^*$ with $y \preceq C^* - \{y\}$. By axiom (ii), there is a finite subset $\{x_1, \ldots, x_n\} \subseteq C^* - \{y\}$ with $y \preceq \{x_1, \ldots, x_n\} - \{y\}$. Now there is $C_0 \in \mathcal{C}$ with $y \in C_0$, and, for each i, there is $C_i \in \mathcal{C}$ with $x_i \in C_i$. Since $\mathcal{C}$ is a chain, one of these, call it C', contains all the others, and the dependent set $\{y, x_1, \ldots, x_n\}$ is contained in C'. But since C' is independent, so are its subsets, and this is a contradiction. Zorn's lemma now provides a maximal element M of X; that is, M is a maximal independent subset of Ω containing B. If M is not a basis, then there exists $x \in \Omega$ with $x \not\preceq M$. By Lemma 6.70, $M \cup \{x\}$ is an independent set strictly larger than M, contradicting the maximality of M. Therefore, bases exist. •

Theorem 6.72. *If Ω is a set with a dependency relation $\preceq$, then any two bases B and C have the same cardinality.*

Proof. If $B = \varnothing$, we claim that $C = \varnothing$. Otherwise, there exists $y \in C$ and, since C is independent, $y \not\preceq C - \{y\}$. But $y \preceq B = \varnothing$ and $\varnothing \subseteq C - \{y\}$, so that axiom (iii) gives $y \preceq C - \{y\}$, a contradiction. Therefore, we may assume that both B and C are nonempty.

Now assume that B is finite; say, $B = \{x_1, \ldots, x_n\}$. We prove, by induction on $k \geq 0$, that there exists $\{y_1, \ldots, y_{k-1}\} \subseteq C$ with

$$B_k = \{y_1, \ldots, y_{k-1}, x_k, \ldots, x_n\}$$

a basis: The elements $x_1 \ldots, x_{k-1}$ in B can be replaced by elements $y_1, \ldots, y_{k-1} \in C$ so that B_k is a basis. We define $B_0 = B$, and we interpret the base step to mean that if none of the elements of B are replaced, then $B = B_0$ is a basis; this is obviously true. For the inductive step, assume that $B_k = \{y_1, \ldots, y_{k-1}, x_k, \ldots, x_n\}$ is a basis. We claim that there is $y \in C$ with $y \not\preceq B_k - \{x_k\}$. Otherwise, $y \preceq B_k - \{x_k\}$ for all $y \in C$. But $x_k \preceq C$, because C is a basis, and so axiom (iii) gives $x_k \preceq B_k - \{x_k\}$, contradicting the independence of B_k. Hence, we may choose $y_k \in C$ with $y_k \not\preceq B_k - \{x_k\}$. By Lemma 6.70, the set B_{k+1}, defined by

$$B_{k+1} = (B_k - \{x_k\}) \cup \{y_k\} = \{y_1, \ldots, y_k, x_{k+1}, \ldots, x_n\},$$

is independent. To see that B_{k+1} is a basis, it suffices to show that it generates Ω. Now $y_k \preceq B_k$ (because B_k is a basis), and $y_k \not\preceq B_k - \{x_k\}$; the exchange axiom gives $x_k \preceq (B_k - \{x_k\}) \cup \{y_k\} = B_{k+1}$. By axiom (i), all the other elements of B_k are dependent on B_{k+1}. Now each element of Ω is dependent on B_k, and each element of B_k is dependent on B_{k+1}. By axiom (iii), B_{k+1} generates Ω.

If $|C| > n = |B|$, that is, if there are more y's than x's, then $B_n \subsetneq C$. Thus a proper subset of C generates Ω, and this contradicts the independence of C. Therefore, $|C| \leq |B|$. It follows that C is finite, and so the preceding argument can be repeated, interchanging the roles of B and C. Hence, $|B| \leq |C|$, and we conclude that $|B| = |C|$ if Ω has a finite basis.

When B is infinite, the reader may complete the proof by adapting the proof of Theorem 6.51. In particular, replace $\text{supp}(v)$ in that proof by axiom (ii) in the definition of dependency relation. •

We now apply this general result to algebraic dependence.

Definition. If E/k is a field extension, then a ***transcendence basis*** is a maximal algebraically independent subset of E over k, and the ***transcendence degree*** of E/k is defined by

$$\text{tr. deg}(E/k) = |B|.$$

The next theorem shows that transcendence degree is well-defined.

Theorem 6.73. *If E/k is a field extension, then there exists a transcendence basis B. If $F = k(B)$, then F/k is purely transcendental and E/F is algebraic. Moreover, if B and C are maximal algebraically independent subsets, then $|B| = |C|$.*

Proof. In Lemma 6.69, we saw that $\alpha \preceq S$, defined by α being algebraic over $k(S)$, is a dependency relation. By Theorems 6.71 and 6.72, transcendence bases exist, and any two of them have the same cardinality; that is, transcendence degree is well-defined. It remains to show that if B is a transcendence basis, then $E/k(B)$ is algebraic. If not, then there exists $\alpha \in E$ with α transcendental over $k(B)$. By Lemma 6.70, $B \cup \{\alpha\}$ is algebraically independent, and this contradicts the maximality of B. •

Example 6.74.

(i) Intermediate fields F, as in the statement of Theorem 6.73, need not be unique. For example, if $E = \mathbb{Q}(\pi)$, then $\mathbb{Q}(\pi^4)$ and $\mathbb{Q}(\pi^2)$ are such intermediate fields.

(ii) If $E = k(x_1, \ldots, x_n)$ is the field of rational functions in n variables over a field k, then $\text{tr. deg}(E/k) = n$, for $\{x_1, \ldots, x_n\}$ is a transcendence basis of E.

(iii) If E/k is a field extension, then E/k is algebraic if and only if $\text{tr. deg}(E/k) = 0$. ◂

Here is a small application of transcendence degree.

Proposition 6.75. *There are nonisomorphic fields each of which is isomorphic to a subfield of the other.*

Proof. Clearly, $\mathbb{C}$ is isomorphic to a subfield of $\mathbb{C}(x)$. However, we claim that $\mathbb{C}(x)$ is isomorphic to a subfield of $\mathbb{C}$. Let B be a transcendence basis of $\mathbb{C}$ over $\mathbb{Q}$, and discard one of its elements, say, b. The algebraic closure F of $\mathbb{Q}(B - \{b\})$ is a proper subfield of $\mathbb{C}$, for $b \notin F$; in fact, b is transcendental over F, by Proposition 6.67. Therefore, $F \cong \mathbb{C}$, by Exercise 6.54 on page 375, and so $F(b) \cong \mathbb{C}(x)$. Therefore, each of $\mathbb{C}$ and $\mathbb{C}(x)$ is isomorphic to a subfield of the other. On the other hand $\mathbb{C}(x) \not\cong \mathbb{C}$, because $\mathbb{C}(x)$ is not algebraically closed. •

We continue our investigation into the structure of fields by considering separability in more detail. Recall that if E/k is a field extension, then an element $\alpha \in E$ is *separable*

over k if either α is transcendental over k or $\operatorname{irr}(\alpha, k)$ is a separable polynomial[9]; that is, $\operatorname{irr}(\alpha, k)$ has no repeated roots. An extension E/k is *separable* if every $\alpha \in E$ is separable over k; otherwise, it is *inseparable*.

Proposition 6.76. *Let $f(x) \in k[x]$, where k is a field, and let $f'(x)$ be its derivative.*

(i) *$f(x)$ has repeated roots if and only if $(f, f') \neq 1$.*

(ii) *If k is a field of characteristic $p > 0$, then $f'(x) = 0$ if and only if $f(x) \in k[x^p]$.*

(iii) *If k is a field of characteristic $p > 0$ and if $f'(x) = 0$, then $f(x)$ has no repeated roots. Conversely, if $f(x)$ is an irreducible polynomial in $k[x]$, then the conditions in parts* (i) *and* (ii) *are all equivalent.*

Proof. (i) If $f(x)$ has repeated roots, then $f(x) = (x-\alpha)^2 g(x)$ in $k[x]$, so that $f'(x) = 2(x-\alpha)g(x) + (x-\alpha)^2 g'(x)$. Therefore, $x - \alpha$ is a common divisor of $f(x)$ and $f'(x)$, and so $(f, f') \neq 1$.

Conversely, it suffices to work in a splitting field of $f(x)$, by Corollary 3.41. If $x - \alpha$ is a divisor of (f, f'), then $f(x) = (x-\alpha)u(x)$ and $f'(x) = (x-\alpha)v(x)$. The product rule gives $f'(x) = u(x) + (x-\alpha)u'(x)$, so that $u(x) = (x-\alpha)(v(x) - u'(x))$. Therefore,

$$f(x) = (x-\alpha)u(x) = (x-\alpha)^2(v(x) - u'(x)),$$

and so $f(x)$ has a repeated root.

(ii) Assume that $f(x) = \sum_i a_i x^i$ and $f'(x) = 0 = \sum_i i a_i x^{i-1}$. If the coefficient $a_i \neq 0$, then $i a_i x^{i-1} = 0$ if and only if $i a_i = 0$; this happens only if $p \mid i$. Therefore, the only nonzero coefficents of $f(x)$ must be of the form a_i for $p \mid i$; that is, $f(x) \in k[x^p]$.

If $f(x) \in k[x^p]$, then $f(x) = \sum_j a_{pj} x^{pj}$ and $f'(x) = \sum_j pj a_{pj} x^{pj-1} = 0$.

(iii) If $f'(x) = 0$, then $(f, f') = (f, 0) = f$; hence, if $f(x)$ is not constant [in particular, if $f(x)$ is irreducible], then $(f, f') \neq 1$.

Conversely, if $f(x)$ is irreducible, then $(f, f') = 1$ or $(f, f') = f$. Now $(f, f') \neq 1$, so that $(f, f') = f$ and, hence, $f \mid f'$. We claim that $f'(x) = 0$. If, on the contrary, $f'(x) \neq 0$, then $f'(x)$ has a degree and $\deg(f') < \deg(f)$. But $f \mid f'$ implies $\deg(f) \leq \deg(f')$, and this is a contradiction. Hence, $f'(x) = 0$. •

Corollary 6.77. *If k is a field of characteristic $p > 0$ and $f(x) \in k[x]$, then there exists $e \geq 0$ and a polynomial $g(x) \in k[x]$ with $g(x) \notin k[x^p]$ and $f(x) = g(x^{p^e})$. Moreover, if $f(x)$ is irreducible, then $g(x)$ is separable.*

Proof. If $f(x) \notin k[x^p]$, define $g(x) = f(x)$; if $f(x) \in k[x^p]$, there is $f_1(x) \in [x]$ with $f(x) = f_1(x^p)$. Note that $\deg(f) = p \deg(f_1)$. If $f_1(x) \notin k[x^p]$, define $g(x) = f_1(x)$; otherwise, there is $f_2(x) \in k[x]$ with $f_1(x) = f_2(x^p)$; that is,

$$f(x) = f_1(x^p) = f_2(x^{p^2}).$$

[9]Recall that an irreducible polynomial is *separable* if it has no repeated roots, and an arbitrary polynomial is *separable* if each of its irreducible factors has no repeated roots.

Since $\deg(f) > \deg(f_1) > \cdots$, iteration of this procedure must end after a finite number e of steps. Thus, $f(x) = g(x^{p^e})$, where $g(x)$, defined by $g(x) = f_e(x)$, does not lie in $k[x^p]$. If, now, $f(x)$ is irreducible, then $f_1(x)$ is irreducible, for a factorization of $f_1(x)$ would give a factorization of $f(x)$. It follows that $f_i(x)$ is irreducible for all i. In particular, $f_e(x)$ is irreducible, and so it is separable, by Proposition 6.76(iii). •

Definition. Let k be a field of characteristic $p > 0$, and let $f(x) \in k[x]$. If $f(x) = g(x^{p^e})$, where $g(x) \in k[x]$ but $g(x) \notin k[x^p]$, then

$$\deg(f) = p^e \deg(g).$$

We call p^e the ***degree of inseparability*** of $f(x)$, and we call $\deg(g)$ the ***reduced degree*** of $f(x)$.

Example 6.78.
Let $f(x) = x^{p^3} + x^p + t \in \mathbb{F}_p(t)[x]$. If $g(x) = x^{p^2} + x + t$, then $g(x)$ is separable (for $g'(x) = 1 \neq 0$). Therefore, $f(x)$ has degree of inseparability p and reduced degree p^2. ◂

If k is a field of prime characteric $p > 0$, then the Frobenius map $F : k \to k$, defined by $F : \alpha \mapsto \alpha^p$, is a homomorphism [because $(\alpha + \beta)^p = \alpha^p + \beta^p$]. As any homomorphism of fields, F is an injection. Denote im F by k^p, so that k^p is the subfield of k consisting of all the pth powers of elements in k:

$$k^p = \operatorname{im} F = \{a^p : a \in k\}.$$

To say that F is surjective, that is, $k = k^p$, is to say that every element in k has a pth root in k.

Definition. A field k is called ***perfect*** if either k has characteristic 0 or if k has characteristic $p > 0$ and $k = k^p$.

Existence of pth roots in k is closely related to separability.

Proposition 6.79.

(i) *A field k is perfect if and only if every polynomial in $k[x]$ is separable.*

(ii) *If k is a perfect field, then every algebraic extension E/k is a separable extension.*

(iii) *Every finite field k is perfect, and every algebraic extension E/k is separable. In particular, if $\overline{\mathbb{F}}_p$ is the algebraic closure of $\mathbb{F}_p$, then $\overline{\mathbb{F}}_p/\mathbb{F}_p$ is a separable extension.*

Proof. (i) If k has characteristic 0, then Lemma 4.4 shows that every polynomial in $k[x]$ is separable. Assume now that k has characteristic $p > 0$ and that $f(x) \in k[x]$ is inseparable. By Proposition 6.76, $f(x) \in k[x^p]$, so that $f(x) = \sum_i a_i x^{pi}$. If every element in k has a pth root, then $a_i = b_i^p$ for $b_i \in k$. Hence,

$$f(x) = \sum_i b_i^p x^{pi} = \Big(\sum_i b_i x^i\Big)^p,$$

and so $f(x)$ is not irreducible. In other words, if $k = k^p$, then every irreducible polynomial in $k[x]$ is separable and, hence, every polynomial is separable.

Conversely, assume that every polynomial in $k[x]$ is separable. If k has characteristic 0, there is nothing to prove. If k has characteristic $p > 0$ and if $a \in k$, then $x^p - a$ has repeated roots; since our hypothesis says that irreducible polynomials are separable, $x^p - a$ factors. Proposition 3.126 now says that a has a pth root in k; that is, $a \in k^p$. Therefore, $k = k^p$, and so k is perfect.

(ii) If E/k is an algebraic extension, then every $\alpha \in E$ has a minimum polynomial $\operatorname{irr}(\alpha, k)$; since $\operatorname{irr}(\alpha, k)$ is a separable polynomial, by part (i), α is separable over k, and so E/k is a separable extension.

(iii) As any homomorphism of fields, the Frobenius $F\colon k \to k$ is injective. If k is a finite field of characteristic $p > 0$, then Exercise 1.58 on page 36 shows that F must also be surjective; that is, $k = k^p$. Therefore, k is perfect, and part (ii) gives the rest of the statement. •

We will soon need a variant of Proposition 3.126.

Lemma 6.80. *Let p be a prime, let $e \geq 0$, and let k be a field of characteristic $p > 0$. If $c \in k$ and $c \notin k^p$, then $f(x) = x^{p^e} - c$ is irreducible in $k[x]$.*

Proof. The proof is by induction on $e \geq 0$, the base step being true because every linear polynomial is irreducible. For the inductive step, suppose the statement is false. Let $g(x) \in k[x]$ be irreducible, and let $g(x)^m$, for $m \geq 1$, be the highest power of $g(x)$ dividing $f(x)$:

$$x^{p^e} - c = g(x)^m h(x),$$

where $(g(x), h(x)) = 1$. Take the derivative, $0 = mg(x)^{m-1}g'(x)h(x) + g(x)^m h'(x)$, and divide by $g(x)^{m-1}$,

$$0 = mg'(x)h(x) + g(x)h'(x).$$

Therefore, $h(x) \mid h'(x)$, because $(g, h) = 1$. If $h'(x) \neq 0$, then $\deg(h')$ is defined and $\deg(h') < \deg(h)$, a contradiction; thus, $h'(x) = 0$. Proposition 6.76 gives

$$h(x) = h_1(x^p), \quad \text{where } h_1(x) \in k[x].$$

Now $mg'(x)h(x) = 0$ gives

$$mg'(x) = 0, \tag{5}$$

for $h(x) \neq 0$, and this implies that $(g^m(x))' = 0$. Hence, Proposition 6.76 gives

$$g^m(x) = g_1(x^p), \quad \text{where } g_1(x) \in k[x].$$

Therefore,

$$x^{p^e} - c = g(x)^m h(x) = g_1(x^p)h_1(x^p),$$

and so, replacing x^p by x, we have

$$x^{p^{e-1}} - c = g_1(x)h_1(x).$$

Since $x^{p^{e-1}} - c$ is irreducible, by the inductive hypothesis, one of g_1, h_1 must be constant. But if $g_1(x)$ is constant, then $g_1(x^p)$ is constant and $g^m(x)$ is constant, a contradiction. Therefore, $h_1(x)$ is constant; absorbing it into $g_1(x)$, we have $x^{p^{e-1}} - c = g_1(x)$ and

$$x^{p^e} - c = g_1(x^p) = g(x)^m.$$

If $p \mid m$, then $x^{p^e} - c = (g(x)^p)^{m/p}$, and so all the coefficients lie in k^p, contradicting $c \notin k^p$; therefore, $p \nmid m$. Eq. (5) now gives $g'(x) = 0$, so that $g(x) \in k[x^p]$; say, $g(x) = g_2(x^p)$. This forces $m = 1$, because $x^{p^e} - c = g(x)^m$ gives $x^{p^{e-1}} - c = g_2(x)^m$, which is a forbidden factorization of the irreducible $x^{p^{e-1}} - c$. •

If E/k is a field extension, where k has characteristic p, then $k^p \subseteq E^p$, but we do not know whether $k \subseteq E^p$; that is, E^p may not be an intermediate field of E/k (for example, take $E = k$). Denote the subfield of E obtained by adjoining E^p to k by $k(E^p)$.

Proposition 6.81.

(i) *Let $k \subseteq B \subseteq E$ be a tower of fields with E/k algebraic. If E/k is separable, then E/B is separable.*

(ii) *Let E/k be an algebraic field extension, where k has characteristic $p > 0$. If E/k is a separable extension, then $E = k(E^p)$. Conversely, if E/k is finite and $E = k(E^p)$, then E/k is separable.*

Proof. (i) If $\alpha \in E$, then α is algebraic over B, and $\operatorname{irr}(\alpha, B) \mid \operatorname{irr}(\alpha, k)$ in $B[x]$, for their gcd is not 1 and $\operatorname{irr}(\alpha, B)$ is irreducible. Since $\operatorname{irr}(\alpha, k)$ has no repeated roots, $\operatorname{irr}(\alpha, B)$ has no repeated roots, and hence $\operatorname{irr}(\alpha, B)$ is a separable polynomial. Therefore, E/B is a separable extension.

(ii) Let E/k be a separable extension. Now $k(E^p) \subseteq E$, and so $E/k(E^p)$ is a separable extension, by part (i). But if $\beta \in E$, then $\beta^p \in E^p \subseteq k(E^p)$; say, $\beta^p = \alpha$. Hence, $\operatorname{irr}(\beta, k(E^p)) \mid (x^p - \alpha)$ in $\big(k(E^p)\big)[x]$, and so this polynomial is not separable because it divides $x^p - \alpha = (x - \beta)^p$. We conclude that $\beta \in k(E^p)$; that is, $E = k(E^p)$.

Conversely, suppose that $E = k(E^p)$. We begin by showing that if $\beta_1, \ldots, \beta_s$ is a linearly independent list in E (where E is now viewed only as a vector space over k), then $\beta_1^p, \ldots, \beta_s^p$ is also linearly independent over k. Extend $\beta_1, \ldots, \beta_s$ to a basis $\beta_1, \ldots, \beta_n$ of E, where $n = [E : k]$. Now $\beta_1^p, \ldots, \beta_n^p$ spans E^p over k^p, for if $\eta \in E$, then $\eta = \sum_i a_i \beta_i$, where $a_i \in k$, and hence $\eta^p = \sum_i a_i^p \beta_i^p$. Now take any element $\gamma \in E$. Since $E = k(E^p)$, we have $\gamma = \sum_j c_j \eta_j$, where $c_j \in k$ and $\eta_j \in E^p$. But $\eta_j = \sum_i a_{ji}^p \beta_i^p$ for $a_{ji} \in k$, as we have just seen, so that $\gamma = \sum_i \left(\sum_j c_j a_{ji}^p\right) \beta_i^p$; that is, $\beta_1^p, \ldots, \beta_n^p$ spans E over k. Since $\dim_k(E) = n$, this list is a basis, and hence its sublist $\beta_1^p, \ldots, \beta_s^p$ must be linearly independent over k.

Since E/k is finite, each α is algebraic over k. If $\mathrm{irr}(\alpha, k)$ has degree m, then $1, \alpha, \alpha^2, \ldots, \alpha^m$ is linearly dependent over k, while $1, \alpha, \alpha^2, \ldots, \alpha^{m-1}$ is linearly independent. If α is inseparable, then $\mathrm{irr}(\alpha, k) = f_e(x^{p^e})$ and $m = p^e r$ where r is the reduced degree of $\mathrm{irr}(\alpha, k)$. Since $r = m/p^e < m$, we have $1, \alpha, \alpha^2, \ldots, \alpha^r$ linearly independent over k. But α^{p^e} is a root of $f_e(x)$, so there is a nontrivial dependency relation on $1, \alpha^{p^e}, \alpha^{2p^e}, \ldots, \alpha^{rp^e}$ (for $rp^e = m$). We have seen, in the preceding paragraph, that linear independence of $1, \alpha, \alpha^2, \ldots, \alpha^r$ implies linear independence of $1, \alpha^{p^e}, \alpha^{2p^e}, \ldots, \alpha^{rp^e}$. This contradiction shows that α must be separable over k. •

Corollary 6.82. *Let E/k be a finite separable extension, where k is a field of characteristic p. If a list $\beta_1, \ldots, \beta_r$ in E is linearly independent over k, then for all $e \geq 1$, the list $\beta_1^{p^e}, \ldots, \beta_r^{p^e}$ is also linearly independent over k.*

Proof. The proof is by induction on $e \geq 1$, with the hypothesis of separability used in the form $E = k(E^p)$, as in the proof of Proposition 6.81(ii). •

Corollary 6.83. *If $k \subseteq B \subseteq E$ is a tower of algebraic extensions, then B/k and E/B are separable extensions if and only if E/k is a separable extension.*

Proof. Since B/k and E/B are separable, Proposition 6.81(ii) gives $B = k(B^p)$ and $E = B(E^p)$. Therefore,

$$E = B(E^p) = k(B^p)(E^p) = k(B^p \cup E^p) = k(E^p) \subseteq E,$$

because $B^p \subseteq E^p$. Therefore, E/k is separable, by Proposition 6.81(ii).

Conversely, if every element of E is separable over k, we have, in particular, that each element of B is separable over k; hence, B/k is a separable extension. Finally, Proposition 6.81(i) shows that E/B is a separable extension. •

Proposition 6.84. *If E/K is an algebraic extension, define*

$$E_s = \{\alpha \in E : \alpha \text{ is separable over } k\};$$

then E_s is an intermediate field that is the unique maximal separable extension of k contained in E.

Proof. This follows from Proposition 4.38(ii), for if α, β are separable over k, then $k(\alpha, \beta)/k$ is separable, and hence $\alpha + \beta$, $\alpha\beta$, and α^{-1} are all separable over k. •

Not surprisingly, if E/k is an algebraic extension, then the extension E/E_s has a special property. Of course, E_s is of interest only when k has characteristic $p > 0$ (otherwise, $E_s = E$).

The next type of extension is "complementary" to separable extensions.

Definition. Let E/k be a field extension, where k has characteristic $p > 0$. Then E/k is a ***purely inseparable extension*** if E/k is algebraic and, for every $\alpha \in E$, there is $e \geq 0$ with $\alpha^{p^e} \in k$.

If E/k is a purely inseparable extension and B is an intermediate field, then it is clear that E/B is purely inseparable.

Proposition 6.85. *If E/k is an algebraic field extension, where k has characteristic $p > 0$, then E/E_s is a purely inseparable extension; moreover, if $\alpha \in E$, then* $\operatorname{irr}(\alpha, E_s) = x^{p^m} - c$ *for some $m \geq 0$.*

Proof. If $\alpha \in E$, write $\operatorname{irr}(\alpha, k) = f_e(x^{p^e})$, where $e \geq 0$ and $f_e(x) \in k[x]$ is a separable polynomial. It follows that α^{p^e} is separable over k and $\alpha^{p^e} \in E_s$. If $\alpha \notin E_s$, choose m minimal with $\alpha^{p^m} \in E_s$. Now α is a root of $x^{p^m} - \alpha^{p^m}$, which is irreducible, by Lemma 6.80, and so $\operatorname{irr}(\alpha, E_s) = x^{p^m} - c$, where $c = \alpha^{p^m}$. •

Definition. If E/k is a finite extension, define the ***separability degree*** by $[E : k]_s = [E_s : k]$, and define the ***inseparability degree*** by $[E : k]_i = [E : E_s]$.

Note that E/k is separable if and only if $[E : k]_i = 1$. It is clear that

$$[E : k] = [E : k]_s [E : k]_i.$$

Proposition 6.86. *Let E/k be a finite extension, where k is a field of characteristic $p > 0$. If E/k is purely inseparable, then $[E : k] = p^e$ for some $e \geq 0$. Hence, for some $e \geq 0$,*

$$[E : k]_i = [E : E_s] = p^e.$$

Proof. If $\alpha \in E$, then α is purely inseparable over k; if α is not constant, then $\operatorname{irr}(\alpha, E_s) = x^{p^m} - c$ for some $c \in k$, where $m \geq 1$. Therefore,

$$[E : k] = [E : k(\alpha)][k(\alpha) : k] = [E : k(\alpha)]p^m.$$

Now $[E : k(\alpha)] < [E : k]$; since $E/k(\alpha)$ is purely inseparable, the proof can be completed by induction. The second statement follows from Proposition 6.85, for E is purely inseparable over E_s. •

Proposition 6.87. *If $k \subseteq B \subseteq E$ is a tower of finite extensions, where k is a field of characteristic $p > 0$, then*

$$[E : k]_s = [E : B]_s [B : k]_s \quad \textit{and} \quad [E : k]_i = [E : B]_i [B : k]_i.$$

Proof. In light of the equation $[E : k] = [E : k]_s [E : k]_i$, it suffices to prove $[E : k]_s = [E : B]_s [B : k]_s$.

The notation B_s is unambiguous, but the notation E_s here is ambiguous. We write E_s to denote the intermediate field consisting of all those elements of E that are separable over k, and we write

$$E_B = \{\alpha \in E : \alpha \text{ is separable over } B\}.$$

We have $k \subseteq B_s \subseteq E_s \subseteq E_B \subseteq E$; let us see that $E_s \subseteq E_B$. If $\alpha \in E$ is separable over k, then $\operatorname{irr}(\alpha, k)$ has no repeated roots; hence, α is separable over B, because $\operatorname{irr}(\alpha, B) \mid \operatorname{irr}(\alpha, k)$ in $B[x]$, and so $\alpha \in E_B$. With this notation,

$$[E : k]_s = [E_s : k], \quad [E : B]_s = [E_B : B], \quad \text{and} \quad [B : k]_s = [B_s : k].$$

Now

$$[E : k]_s = [E_s : k] = [E_s : B_s][B_s : k] = [E_s : B_s][B : k]_s.$$

Thus, it suffices to prove

$$[E_s : B_s] = [E_B : B],$$

for $[E_B : B] = [E : B]_s$.

We show that $[E_s : B_s] \le [E_B : B]$ by proving that a list $\beta_1, \ldots, \beta_r$ in $E_s \subseteq E_B$ linearly independent over B_s is also linearly independent over B. Suppose that $\sum b_i \beta_i = 0$, where $b_i \in B$ are not all 0. For all $e \ge 0$, we have $0 = (\sum b_i \beta_i)^{p^e} = \sum b_i^{p^e} \beta_i^{p^e}$. But there is $e \ge 0$ with $b_i^{p^e} \in B_s$ for all i, because B/B_s is purely inseparable, and so the list $\beta_1^{p^e}, \ldots, \beta_r^{p^e}$ is linearly dependent over B_s, contradicting Corollary 6.82 (for E_s/B_s is a separable extension). For the reverse inequality $[E_s : B_s] \ge [E_B : B]$, take a list $\gamma_1, \ldots, \gamma_t$ in E_B that is linearly independent over B. Since E_B/E_s is purely inseparable (it is an intermediate field of E/E_s), there is $e \ge 0$ with $\gamma_i^{p^e} \in E_s$ for all i. But E_s/B is a separable extension, so that Corollary 6.82 gives $\gamma_1^{p^e}, \ldots, \gamma_t^{p^e}$ linearly independent over B; a fortiori, $\gamma_1^{p^e}, \ldots, \gamma_t^{p^e}$ is linearly independent over B_s. Therefore, $[E_s : B_s] = [E_B : B]$. •

We merely state some further results about separability.

Definition. If A and B are intermediate fields of a field extension E/k, then A and B are ***linearly disjoint*** if every finite list $\alpha_1, \ldots, \alpha_n$ in A that is linearly independent over k is linearly independent over B. That is, if $\sum_i c_i \alpha_i = 0$ implies all $c_i = 0$ whenever $c_i \in k$, then $\sum_i \beta_i \alpha_i = 0$ implies all $\beta_i = 0$ whenever $\beta_i \in B$.

This condition on A and B can be shown to be symmetric; that is, every finite list in B that is linearly independent over k is also linearly independent over A. In Chapter 4, we defined two intermediate fields A and B to be *linearly disjoint* if $A \cap B = k$. This new definition is stronger than the old one: If $\alpha \in A$ and $\alpha \notin k$, then $1, \alpha$ is linearly independent over k. If $\alpha \in A \cap B$, then $-\alpha \cdot 1 + 1 \cdot \alpha = 0$ is a dependency relation over B (for $-\alpha, 1 \in B$). However, there are examples of intermediate fields A and B with $A \cap B = k$ that are not linearly disjoint in this new sense.

Definition. Let k be a field of characteristic p; for $n \geq 1$, define

$$k^{1/p} = \{\alpha \in \overline{k} : \alpha^p \in k\},$$

where $\overline{k}$ is the algebraic closure of k.

Theorem. *An algebraic field extension E/k is separable if and only if $k^{1/p}$ and E are linearly disjoint (as intermediate fields of $\overline{E}/k$, where $\overline{E}$ is the algebraic closure of E).*

Proof. See Zariski–Samuel, *Commutative Algebra* I, page 109. •

If we do not assume that a field extension E/k is algebraic, are the generalizations of Propositions 6.81(ii) through 6.85 still true?

Definition. A ***separating transcendence basis*** of a field extension E/k is a transcendence basis B with $E/k(B)$ a separable extension.

Not every extension E/k has a separating transcendence basis. For example, if E/k is an inseparable algebraic extension, then the only transcendence basis is $\varnothing$; but $k(\varnothing) = k$, and $E/k(\varnothing)$ is inseparable.

Theorem (Mac Lane). *If a field extension E/k has a separating transcendence basis, then E and $k^{1/p}$ are linearly disjoint intermediate fields of $\overline{E}$, the algebraic closure of E. Conversely, if E and $k^{1/p}$ are linearly disjoint and E/k is finitely generated, that is, $E = k(u_1, \ldots, u_n)$, then E/k has a separating transcendence basis.*

Proof. See Jacobson, *Basic Algebra* II, page 519. •

The following example shows why one assumes, in Mac Lane's theorem, that E/k is finitely generated.

Example 6.88.
Let k be a perfect field of characteristic p, let $k(x)$ be the function field, and define

$$E = k(\{u_n, \text{ for } n \geq 1 : u_n^{p^n} = x\}).$$

Since k is perfect, every extension of k is separable, and so $E \cap k^{1/p} = k$. However, we claim that E/k does not have a separating transcendence basis. By Exercise 6.52 on page 375, tr. deg$(E/k) = 1$, because any pair x^{1/p^n} and x^{1/p^m} are algebraically dependent; let $\{\beta\}$ be a transcendence basis. Now $k(\beta) \neq E$, and so there exists some u_n with $u_n \notin k(\beta)$; choose n minimal. Consider the tower $k(\beta) \subseteq k(\beta, u_n) \subseteq E$. If $\{\beta\}$ were a separating transcendence basis, then $E/k(\beta, u_n)$ would be separable, by Proposition 6.81(i). But $\text{irr}(u_n, k(\beta))$ is a nonlinear divisor of $y^{p^n} - x^{p^n}$, because $u_n \notin k(\beta)$, and hence it has repeated roots; therefore, $E/k(\beta, u_n)$ is inseparable, a contradiction. ◂

EXERCISES

6.41 Let k be a field of characteristic $p > 0$, and let $f(x) = x^{2p} - x^p + t \in k(t)[x]$.

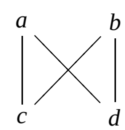
(i) Prove that $f(x)$ is an irreducible polynomial in $k(t)[x]$.

(ii) Prove that $f(x)$ is inseparable.

(iii) Prove that there exists an algebraic extension $E/k(t)$ for which there is no intermediate field E_i with E_i/k purely inseparable and E/E_i separable. (Compare with Corollary 6.85 and Proposition 4.38.)

6.42 Let m be a positive integer, and let X be the set of all its (positive) divisors. Prove that X is a partially ordered set if one defines $a \preceq b$ to mean $a \mid b$.

6.43 Recall that if S is a subset of a partially ordered set X, then the ***least upper bound*** of S (should it exist) is an upper bound m of S such that $m \preceq u$ for every upper bound u of S. If X is the following partially ordered set (in which $d \preceq a$ is indicated by a joining with a line and having a higher than d),



prove that the subset $S = \{c, d\}$ has an upper bound but no least upper bound.

6.44 Let G be an abelian group, and let $S \subseteq G$ be a subgroup.

(i) Prove that there exists a subgroup H of G maximal with the property that $H \cap S = \{0\}$. Is this true if G is not abelian?

(ii) If H is maximal with $H \cap S = \{0\}$, prove that $G/(H + S)$ is torsion.

6.45 Call a subset C of a partially ordered set X ***cofinal*** if, for each $x \in X$, there exists $c \in C$ with $x \preceq c$.

(i) Prove that $\mathbb{Q}$ and $\mathbb{Z}$ are cofinal subsets of $\mathbb{R}$.

(ii) Prove that every chain X contains a well-ordered cofinal subset.

Hint. Use Zorn's lemma on the family of all the well-ordered subsets of X.

(iii) Prove that every well-ordered subset in X has an upper bound if and only if every chain in X has an upper bound.

6.46 (i) Give an example of a commutative ring containing two prime ideals P and Q for which $P \cap Q$ is not a prime ideal.

(ii) If $P_1 \supseteq P_2 \supseteq \cdots P_n \supseteq P_{n+1} \cdots$ is a decreasing sequence of prime ideals in a commutative ring R, prove that $\bigcap_{n \geq 1} P_n$ is a prime ideal.

(iii) Prove that every commutative ring R has a ***minimal prime ideal***; that is, a prime ideal I for which there is no prime ideal P with $P \subsetneq I$.

Hint. Partially order the set of all prime ideals by reverse inclusion: $P \preceq Q$ means $P \supseteq Q$.

6.47 Let V be a vector space, and let S be a subspace of V. Prove that there exists a subspace W of V maximal with the property that $W \cap S = 0$ and that $V = S \oplus W$.

6.48 Recall that a subset S of a commutative ring R is called ***multiplicatively closed*** if $0 \notin S$ and $s, s' \in S$ implies $ss' \in S$. Complete Exercise 6.9 on page 325 by proving that if S is

a multiplicatively closed set with $S \cap I = \varnothing$, then there exists an ideal J maximal with the property that J contains I and $J \cap S = \varnothing$.

6.49 Prove that every nonunit in a commutative ring lies in some maximal ideal. [This result was used to solve Exercise 6.16(ii) on page 326.]

6.50 If $p_1, \ldots, p_n$ are distinct primes in $\mathbb{Z}$, prove that $\sqrt{p_1}, \ldots, \sqrt{p_n}$ is a linearly independent list over $\mathbb{Q}$.

6.51 Prove that a field extension E/k may not have an intermediate field K with K/k algebraic and E/K purely transcendental.
Hint. Prove that there is no intermediate field K with $\mathbb{Q} \subseteq K \subsetneq \mathbb{C}$ with $\mathbb{C}/K$ purely transcendental.

6.52 If $E = k(X)$ is an extension of a field k, and if every pair $u, v \in X$ is algebraically dependent, prove that tr. $\deg(E/k) \le 1$. Conclude that if

$$k \subseteq k_1 \subseteq k_2 \subseteq \cdots$$

is a tower of fields with tr. $\deg(k_n/k) = 1$ for all $n \ge 1$, then tr. $\deg(k^*/k) = 1$, where $k^* = \bigcup_{n \ge 1} k_n$.

6.53 Prove that if k is the prime field of a field E and if tr. $\deg(E/k) \le \aleph_0$, then E is countable.

6.54 Prove that two algebraically closed fields of the same characteristic are isomorphic if and only if they have the same transcendence degree over their prime fields.
Hint. Use Lemma 6.61.

6.55 (i) If $k \subseteq B \subseteq E$ is a tower of fields, prove that

$$\text{tr. deg}(E/k) = \text{tr. deg}(E/B) + \text{tr. deg}(B/k).$$

Hint. Prove that if X is a transcendence basis of B/k and Y is a transcendence basis of E/B, then $X \cup Y$ is a transcendence basis for E/k.

(ii) Let E/k be a field extension, and let B and C be intermediate fields. Prove that

$$\text{tr. deg}(B \vee C) + \text{tr. deg}(B \cap C) = \text{tr. deg}(B) + \text{tr. deg}((C),$$

where $B \vee C$ is the compositum.

Hint. Extend a transcendence basis of $B \cap C$ to a transcendence basis of B and to a transcendence basis of C.

6.56 Prove that $\varphi \in k(x)$ has degree 1 if and only if φ is a linear fractional transformation.

6.57 Prove, for any field k, that $\text{PGL}(2, k) \cong \text{LF}(k)$, where $\text{PGL}(2, k) = \text{GL}(2, k)/Z(2, k)$ and $Z(2, k)$ is the (normal) subgroup of $\text{GL}(2, k)$ consisting of all the (nonzero) scalar matrices [$Z(2, k)$ is the center of $\text{GL}(2, k)$].

6.58 Prove that if E/k is an algebraic extension and $\beta \in E$ is both separable and purely inseparable, then $\beta \in k$.

6.59 Give an example of two intermediate fields A and B of an extension E/k with $A \cap B = k$ but that are not linearly disjoint in the sense of the definition on page 372.

6.5 VARIETIES

Analytic geometry gives pictures of equations. For example, we picture a function $f: \mathbb{R} \to \mathbb{R}$ as its graph, which consists of all the ordered pairs $(a, f(a))$ in the plane; that is, f is the set of all the solutions $(a, b) \in \mathbb{R}^2$ of

$$g(x, y) = y - f(x) = 0.$$

We can also picture equations that are not graphs of functions. For example, the set of all the zeros of the polynomial

$$h(x, y) = x^2 + y^2 - 1$$

is the unit circle. We can also picture simultaneous solutions in $\mathbb{R}^2$ of several polynomials of two variables, and, indeed, we can picture simultaneous solutions in $\mathbb{R}^n$ of several polynomials of n variables. But there is a very strong connection between the rings $k[x_1, \ldots, x_n] = k[X]$ and the geometry of subsets of k^n going far beyond this. Given a set of polynomials $f_1(X), \ldots, f_t(X)$ in n variables, call the subset $V \subseteq k^n$ consisting of their common zeros a *variety*. Of course, we can study varieties because solutions of systems of polynomial equations (an obvious generalization of systems of linear equations) are intrinsically interesting. On the other hand, some systems are more interesting than others; investigating a problem may lead to a variety, and understanding the variety and its properties (e.g., irreducibility, dimension, genus, singularities, and so forth) may contribute to an understanding of the original problem. For example, Leibniz raised the question of determining those functions that could be integrated explicitly in terms of "elementary functions:" algebraic combinations of polynomials, trigonometric and inverse trigonometric functions, exponentials, and logarithms. In 1694, John Bernoulli conjectured that the integral arising from the arclength of an ellipse could not be so integrated. Similar integrals arise in finding periods of pendulums, as well as in other problems in mechanics, and all of them can be reduced to the form

$$\int_0^x \frac{dt}{\sqrt{p(t)}},$$

where $p(t)$ is either a cubic or quartic polynomial; that is, the polynomial $y^2 - p(x)$ in $\mathbb{R}[x, y]$ has arisen. In analogy to

$$\sin^{-1} x = \int_0^x \frac{dt}{\sqrt{1 - t^2}},$$

Jacobi introduced the inverse function

$$u^{-1}(x) = \int_0^x \frac{dt}{\sqrt{p(t)}},$$

and he called $u(x)$ an ***elliptic function***. Just as $\sin x$ determines the unit circle via the parametrization $(\sin x, \cos x)$, so, too, does an elliptic function determine a curve via the

parametrization $(u(x), u'(x))$, where $u'(x)$ is the derivative of $u(x)$. It was also noted that elliptic functions are *periodic* (as is $\sin x$); that is, there is some number q with $u(x+mq) = u(x)$ for all real x and all $m \in \mathbb{Z}$. With the development of integration of functions of a complex variable, Gauss viewed elliptic functions as

$$u^{-1}(z) = \int_0^z \frac{dw}{\sqrt{p(w)}},$$

where $p(w)$ is either a cubic or quartic polynomial in $\mathbb{C}[w]$; that is, the polynomial $u^2 - p(z)$ in $\mathbb{C}[z, u]$ has arisen. In viewing elliptic functions in this way, he saw that they are *doubly periodic*; that is, there are (complex) numbers q and r so that

$$u(z + mq + nr) = u(z)$$

for all complex z and all $m, n \in \mathbb{Z}$. Moreover, $u(z)$ determines a complex curve (called an ***elliptic curve***) consisting of all $(u(z), u'(z))$. A one-dimensional complex space is a two-dimensional real space, and double periodicity says that this complex curve is a torus; that is, the surface of a doughnut, possibly having several holes. One consequence is that the behavior of an elliptic function depends on whether the associated curve is nonsingular; that is, whether it has an appropriate tangent space at every point. The subject was further enriched when Riemann introduced Riemann surfaces into the study of elliptic functions and elliptic curves. This is the beginning of a very rich subject[10]; indeed, further deep investigations of such matters were essential in A. Wiles's proof of Fermat's last theorem, in which he proves elliptic curves have certain sophisticated properties. More generally, the interplay between $k[x_1, \ldots, x_n]$ and varieties has evolved into what is nowadays called *algebraic geometry*, and this section may be regarded as an introduction to this subject.

Notation. Let k be a field and let k^n denote the set of all n-tuples

$$k^n = \{a = (a_1, \ldots, a_n) : a_i \in k \text{ for all } i\}.$$

The polynomial ring $k[x_1, \ldots, x_n]$ in several variables may be denoted by $k[X]$, where X is the abbreviation

$$X = (x_1, \ldots, x_n).$$

In particular, $f(X) \in k[X]$ may abbreviate $f(x_1, \ldots, x_n) \in k[x_1, \ldots, x_n]$.

In what follows, we regard polynomials $f(x_1, \ldots, x_n) \in k[x_1, \ldots, x_n]$ as functions of n variables $k^n \to k$. Here is the precise definition.

Definition. If $f(X) \in k[X]$ define its ***polynomial function*** $f^\flat : k^n \to k$ by evaluation: If $(a_1, \ldots, a_n) \in k^n$, then

$$f^\flat : (a_1, \ldots, a_n) \mapsto f(a_1, \ldots, a_n).$$

The next proposition generalizes Corollary 3.28 from one variable to several variables.

[10]For a leisurely and more detailed account of the development of elliptic functions, see Chapters 14 and 15 of Stillwell, *Mathematics and Its History*.

Proposition 6.89. *Let k be an infinite field and let $k[X] = k[x_1, \dots, x_n]$. If $f(X), g(X) \in k[X]$ satisfy $f^\flat = g^\flat$, then $f(x_1, \dots, x_n) = g(x_1, \dots, x_n)$.*

Proof. The proof is by induction on $n \geq 1$; the base step is Corollary 3.28. For the inductive step, write

$$f(X, y) = \sum_i p_i(X)y^i \quad \text{and} \quad g(X, y) = \sum_i q_i(X)y^i,$$

where X denotes $(x_1, \dots, x_n)$. If $f^\flat = g^\flat$, then we have $f(a, \beta) = g(a, \beta)$ for every $a \in k^n$ and every $\beta \in k$. For fixed $a \in k^n$, define $F_a(y) = \sum_i p_i(a)y^i$ and $G_a(y) = \sum_i q_i(a)y^i$. Since both $F_a(y)$ and $G_a(y)$ are in $k[y]$, the base step gives $p_i(a) = q_i(a)$ for all $a \in k^n$. By the inductive hypothesis, $p_i(X) = q_i(X)$ for all i, and hence

$$f(X, y) = \sum_i p_i(X)y^i = \sum_i q_i(X)y^i = g(X, y),$$

as desired. •

As a consequence of this last proposition, we drop the $f^\flat$ notation and identify polynomials with their polynomial functions when k is infinite.

Definition. If $f(X) \in k[X] = k[x_1, \dots, x_n]$ and $f(a) = 0$, where $a \in k^n$, then a is called a ***zero*** of $f(X)$. [If $f(x)$ is a polynomial in one variable, then a zero of $f(x)$ is also called a *root* of $f(x)$.]

Proposition 6.90. *If k is an algebraically closed field and $f(X) \in k[X]$ is not a constant, then $f(X)$ has a zero.*

Proof. We prove the result by induction on $n \geq 1$, where $X = (x_1, \dots, x_n)$. The base step follows at once from our assuming that $k^1 = k$ is algebraically closed. As in the previous proof, write

$$f(X, y) = \sum_i g_i(X)y^i.$$

For each $a \in k^n$, define $f_a(y) = \sum_i g_i(a)y^i$. If $f(X, y)$ has no zeros, then each $f_a(y) \in k[y]$ has no zeros, and the base step says that $f_a(y)$ is a nonzero constant for all $a \in k^n$. Thus, $g_i(a) = 0$ for all $i > 0$ and all $a \in k^n$. By Proposition 6.89, which applies because algebraically closed fields are infinite, $g_i(X) = 0$ for all $i > 0$, and so $f(X, y) = g_0(X)y^0 = g_0(X)$. By the inductive hypothesis, $g_0(X)$ is a nonzero constant, and the proof is complete. •

We now give some general definitions describing solution sets of polynomials.

Definition. If F is a subset of $k[X] = k[x_1, \ldots, x_n]$, then the ***variety***[11,12] defined by F is

$$\text{Var}(F) = \{a \in k^n : f(a) = 0 \text{ for every } f(X) \in F\};$$

thus, $\text{Var}(F)$ consists of all those $a \in k^n$ which are zeros of every $f(X) \in F$.

Example 6.91.
(i) If k is algebraically closed, then Proposition 6.90 says that if $f(X) \in k[X]$ is not constant, then $\text{Var}(f(X)) \neq \varnothing$.

(ii) Here are some varieties defined by two equations:

$$\text{Var}(x, y) = \{(a, b) \in k^2 : x = 0 \text{ and } y = 0\} = \{(0, 0)\}$$

and

$$\text{Var}(xy) = x\text{-axis} \cup y\text{-axis}.$$

(iii) Here is an example in higher-dimensional space. Let A be an $m \times n$ matrix with entries in k. A system of m equations in n unknowns,

$$AX = B,$$

where B is an $n \times 1$ column matrix, defines a variety, $\text{Var}(AX = B)$, which is a subset of k^n. Of course, $AX = B$ is really shorthand for a set of m linear equations in n variables, and $\text{Var}(AX = B)$ is usually called the ***solution set*** of the system $AX = B$; when this system is homogeneous, that is, when $B = 0$, then $\text{Var}(AX = 0)$ is a subspace of k^n, called the ***solution space*** of the system. ◂

The next result shows that, as far as varieties are concerned, we may just as well assume that the subsets F of $k[X]$ are ideals of $k[X]$.

[11]There is some disagreement about the usage of this term. Some call this an *affine variety*, in contrast to the analogous *projective variety*. Some insist that varieties should be *irreducible*, which we will define later in this section.

[12]The term *variety* arose as a translation by E. Beltrami (inspired by Gauss) of the German term *Mannigfaltigkeit* used by Riemann; nowadays, this term is usually translated as *manifold*. The following correspondence, from Aldo Brigaglia to Steven Kleiman, contains more details.

"I believe the usage of the word *varietà* by Italian geometers arose from the (unpublished) Italian translation of Riemann's Habilitationsvortrag, which was later translated into French by J. Hoüel and published in the Italian journal *Annali*. Indeed, Beltrami wrote to Hoüel on 8 January, 1869:

> J'ai traduit *Mannigfaltigkeit* par *varietà*, dans le sens de *multitudo variarum rerum*. . .

And later, on 14 February, 1869, he wrote

> Je croirais toujours convenable de traduire *Mannigfaltigkeit* par *variété*: j'ai remarqué que Gauss, dans ses Mémoires sur les résidus biquadratiques appelle en latin *varietas* la même chose qui, dans les comptes-rendus rédigés par lui même en allemand dans les *Gelehrte Anzeige*, est désignée par *Mannigfaltigkeit*.

The correspondence of Beltrami and Hoüel can be found in the beautiful book *La découverte de la géométrie non euclidienne sur la pseudosphère: les lettres d'Eugenio Beltrami à Jules Hoüel* (1868–1881), edited by L. Boi, L. Giacardi, and R. Tazzioli, and published by Blanchard, Paris, 1998."

Proposition 6.92. *Let k be a field, and let F and G be subsets of $k[X]$.*

(i) *If $F \subseteq G \subseteq k[X]$, then* $\mathrm{Var}(G) \subseteq \mathrm{Var}(F)$.

(ii) *If $F \subseteq k[X]$ and $I = (F)$ is the ideal generated by F, then*

$$\mathrm{Var}(F) = \mathrm{Var}(I).$$

Proof. (i) If $a \in \mathrm{Var}(G)$, then $g(a) = 0$ for all $g(X) \in G$; since $F \subseteq G$, it follows, in particular, that $f(a) = 0$ for all $f(X) \in F$.

(ii) Since $F \subseteq (F) = I$, we have $\mathrm{Var}(I) \subseteq \mathrm{Var}(F)$, by part (i). For the reverse inclusion, let $a \in \mathrm{Var}(F)$, so that $f(a) = 0$ for every $f(X) \in F$. If $g(X) \in I$, then $g(X) = \sum_i r_i(X) f_i(X)$, where $r_i(X) \in k[X]$ and $f_i(X) \in F$; hence, $g(a) = \sum_i r_i(a) f_i(a) = 0$ and $a \in \mathrm{Var}(I)$. •

It follows that not every subset of k^n is a variety. For example, if $n = 1$, then $k[x]$ is a PID. Hence, if F is a subset of $k[x]$, then $(F) = (g(x))$ for some $g(x) \in k[x]$, and so

$$\mathrm{Var}(F) = \mathrm{Var}((F)) = \mathrm{Var}((g(x))) = \mathrm{Var}(g(x)).$$

But if $g(x) \neq 0$, then it has only a finite number of roots, and so $\mathrm{Var}(F)$ is finite. If k is algebraically closed, then it is an infinite field, and so most subsets of $k^1 = k$ are not varieties.

In spite of our wanting to draw pictures in the plane, there is a major defect with $k = \mathbb{R}$: Some polynomials have no zeros. For example, $f(x) = x^2 + 1$ has no real roots, and so $\mathrm{Var}(x^2 + 1) = \varnothing$. More generally, $g(x_1, \ldots, x_n) = x_1^2 + \cdots + x_n^2 + 1$ has no zeros in $\mathbb{R}^n$, and so $\mathrm{Var}(g(X)) = \varnothing$. Since we are dealing with (not necessarily linear) polynomials, it is a natural assumption to want all their zeros available. For polynomials in one variable, this amounts to saying that k is algebraically closed and, in light of Proposition 6.90, we know that $\mathrm{Var}(f(X)) \neq \varnothing$ for every nonconstant $f(X) \in k[X]$ if k is algebraically closed. Of course, varieties are of interest for all fields k, but it makes more sense to consider the simplest case before trying to understand more complicated problems. On the other hand, many of the first results are valid for any field k. Thus, we will state the hypothesis needed for each proposition, but the reader should realize that the most important case is when k is algebraically closed.

Here are some elementary properties of Var.

Proposition 6.93. *Let k be a field.*

(i) $\mathrm{Var}(1) = \varnothing$ *and* $\mathrm{Var}(0) = k^n$, *where* 0 *is the zero polynomial.*

(ii) *If I and J are ideals in $k[X]$, then*

$$\mathrm{Var}(IJ) = \mathrm{Var}(I \cap J) = \mathrm{Var}(I) \cup \mathrm{Var}(J),$$

where $IJ = \left\{\sum_i f_i(X)g_i(X) \colon f_i(X) \in I \text{ and } g_i(X) \in J\right\}$.

(iii) *If $\{I_\ell : \ell \in L\}$ is a family of ideals in $k[X]$, then*

$$\mathrm{Var}\Big(\sum_\ell I_\ell\Big) = \bigcap_\ell \mathrm{Var}(I_\ell),$$

where $\sum_\ell I_\ell$ is the set of all finite sums of the form $r_{\ell_1} + \cdots + r_{\ell_q}$ with $r_{\ell_i} \in I_{\ell_i}$.

Proof. (i) That $\mathrm{Var}(1) = \varnothing$ is clear, for the constant polynomial 1 has no zeros. That $\mathrm{Var}(0) = k^n$ is clear, for every point a is a zero of the zero polynomial.

(ii) Since $IJ \subseteq I \cap J$, it follows that $\mathrm{Var}(IJ) \supseteq \mathrm{Var}(I \cap J)$; since $IJ \subseteq I$, it follows that $\mathrm{Var}(IJ) \supseteq \mathrm{Var}(I)$. Hence,

$$\mathrm{Var}(IJ) \supseteq \mathrm{Var}(I \cap J) \supseteq \mathrm{Var}(I) \cup \mathrm{Var}(J).$$

To complete the proof, it suffices to show that $\mathrm{Var}(IJ) \subseteq \mathrm{Var}(I) \cup \mathrm{Var}(J)$. If $a \notin \mathrm{Var}(I) \cup \mathrm{Var}(J)$, then there exist $f(X) \in I$ and $g(X) \in J$ with $f(a) \neq 0$ and $g(a) \neq 0$. But $f(X)g(X) \in IJ$ and $(fg)(a) = f(a)g(a) \neq 0$, because k is a domain. Therefore, $a \notin \mathrm{Var}(IJ)$, as desired.

(iii) For each ℓ, the inclusion $I_\ell \subseteq \sum_\ell I_\ell$ gives $\mathrm{Var}\big(\sum_\ell I_\ell\big) \subseteq \mathrm{Var}(I_\ell)$, and so

$$\mathrm{Var}\Big(\sum_\ell I_\ell\Big) \subseteq \bigcap_\ell \mathrm{Var}(I_\ell).$$

For the reverse inclusion, if $g(X) \in \sum_\ell I_\ell$, then there are finitely many ℓ with $g(X) = \sum_\ell f_\ell$, where $f_\ell(X) \in I_\ell$. Therefore, if $a \in \bigcap_\ell \mathrm{Var}(I_\ell)$, then $f_\ell(a) = 0$ for all ℓ, and so $g(a) = 0$; that is, $a \in \mathrm{Var}\big(\sum_\ell I_\ell\big)$. •

Definition. A ***topological space*** is a set X together with a family $\mathcal{F}$ of subsets of X, called ***closed sets***,[13] which satisfy the following axioms:

(i) $\varnothing \in \mathcal{F}$ and $X \in \mathcal{F}$;

(ii) if $F_1, F_2 \in \mathcal{F}$, then $F_1 \cup F_2 \in \mathcal{F}$; that is, the union of two closed sets is closed;

(iii) if $\{F_\ell : \ell \in L\} \subseteq \mathcal{F}$, then $\bigcap_\ell F_\ell \in \mathcal{F}$; that is, any intersection of closed sets is also closed.

Proposition 6.93 shows that the family of all varieties are the closed sets that make k^n a topological space. Varieties are called ***Zariski closed sets***, and they are very useful in the deeper study of $k[X]$. The usual way of regarding $\mathbb{R}$ as a topological space has many closed sets; for example, every closed interval is a closed set. In contrast, the only Zariski closed sets in $\mathbb{R}$, aside from $\mathbb{R}$ itself, are finite.

Definition. A ***hypersurface*** in k^n is a subset of the form $\mathrm{Var}(f)$ for some nonconstant $f(X) \in k[X]$.

---

[13] We can also define a topological space by specifying its *open subsets* which are defined as complements of closed sets.

Corollary 6.94. *Every variety* $\mathrm{Var}(I)$ *in* k^n *is the intersection of finitely many hypersurfaces.*

Proof. By the Hilbert basis theorem, there are $f_1(X), \ldots, f_t(X) \in k[X]$ with $I = (f_1, \ldots, f_t) = \sum_i (f_i)$. By Proposition 6.93(iii), we have $\mathrm{Var}(I) = \bigcap_i \mathrm{Var}(f_i)$. •

Given an ideal I in $k[X]$, we have just defined its variety $\mathrm{Var}(I) \subseteq k^n$. We now reverse direction: Given a subset $A \subseteq k^n$, we assign an ideal in $k[X]$ to it; in particular, we assign an ideal to every variety.

Definition. If $A \subseteq k^n$, define its ***coordinate ring*** $k[A]$ to be the commutative ring

$$k[A] = \{ f^\flat | A : f(X) \in k[X]\}$$

under pointwise operations [recall that $f^\flat : k^n \to k$ is the polynomial function arising from $f(X)$].

The polynomial $f(x_1, \ldots, x_n) = x_i \in k[X]$, when regarded as a polynomial function, is defined by

$$x_i : (a_1, \ldots, a_n) \mapsto a_i;$$

that is, x_i picks out the ith coordinate of a point in k^n. The reason for the name coordinate ring is that if $a \in V$, then $(x_1(a), \ldots, x_n(a))$ describes a.

There is an obvious ring homomorphism $\mathrm{res}\colon k[X] \to k[A]$, given by $f(X) \mapsto f^\flat | A$, and the kernel of this restriction map is an ideal in $k[X]$. We will assume, from now on, that all fields k are infinite, and so we will drop the notation $f^\flat$.

Definition. If $A \subseteq k^n$, define

$$\mathrm{Id}(A) = \{ f(X) \in k[X] = k[x_1, \ldots, x_n] : f(a) = 0 \text{ for every } a \in A\}.$$

The Hilbert basis theorem tells us that $\mathrm{Id}(A)$ is always a finitely generated ideal.

Proposition 6.95. *If* $A \subseteq k^n$*, then there is an isomorphism*

$$k[X]/\,\mathrm{Id}(A) \cong k[A],$$

where $k[A]$ *is the coordinate ring of* A*.*

Proof. The restriction map $\mathrm{res}\colon k[X] \to k[A]$ is a surjection with kernel $\mathrm{Id}(A)$, and so the result follows from the first isomorphism theorem. Note that two polynomials agreeing on A lie in the same coset of $\mathrm{Id}(A)$. •

Although the definition of $\mathrm{Var}(F)$ makes sense for any subset F of $k[X]$, it is most interesting when F is an ideal. Similarly, although the definition of $\mathrm{Id}(A)$ makes sense for any subset A of k^n, it is most interesting when A is a variety. After all, varieties are comprised of solutions of (polynomial) equations, which is what we care about.

Proposition 6.96. *Let k be a field.*

(i) $\mathrm{Id}(\varnothing) = k[X]$ *and, if k is infinite,* $\mathrm{Id}(k^n) = \{0\}$.

(ii) *If $A \subseteq B$ are subsets of k^n, then* $\mathrm{Id}(B) \subseteq \mathrm{Id}(A)$.

(iii) *If $\{A_\ell : \ell \in L\}$ is a family of subsets of k^n, then*

$$\mathrm{Id}\Big(\bigcup_\ell A_\ell\Big) = \bigcap_\ell \mathrm{Id}(A_\ell).$$

Proof. (i) By definition, $f(X) \in \mathrm{Id}(A)$ for some subset $A \subseteq k^n$ if and only if $f(a) = 0$ for all $a \in A$; hence, if $f(X) \notin \mathrm{Id}(A)$, then there exists $a \in A$ with $f(a) \neq 0$. In particular, if $A = \varnothing$, every $f(X) \in k[X]$ must lie in $\mathrm{Id}(\varnothing)$, for there are no elements $a \in \varnothing$. Therefore, $\mathrm{Id}(\varnothing) = k[X]$.

If $f(X) \in \mathrm{Id}(k^n)$, then $f^\flat = 0^\flat$, and so $f(X) = 0$, by Proposition 6.89, because k is infinite.

(ii) If $f(X) \in \mathrm{Id}(B)$, then $f(b) = 0$ for all $b \in B$; in particular, $f(a) = 0$ for all $a \in A$, because $A \subseteq B$, and so $f(X) \in \mathrm{Id}(A)$.

(iii) Since $A_\ell \subseteq \bigcup_\ell A_\ell$, we have $\mathrm{Id}(A_\ell) \supseteq \mathrm{Id}\big(\bigcup_\ell A_\ell\big)$ for all ℓ; hence, $\bigcap_\ell \mathrm{Id}(A_\ell) \supseteq \mathrm{Id}\big(\bigcup_\ell A_\ell\big)$. For the reverse inclusion, suppose that $f(X) \in \bigcap_\ell \mathrm{Id}(A_\ell)$; that is, $f(a_\ell) = 0$ for all ℓ and all $a_\ell \in A_\ell$. If $b \in \bigcup_\ell A_\ell$, then $b \in A_\ell$ for some ℓ, and hence $f(b) = 0$; therefore, $f(X) \in \mathrm{Id}\big(\bigcup_\ell A_\ell\big)$. •

We would like to have a formula for $\mathrm{Id}(A \cap B)$. Certainly, it is not true that $\mathrm{Id}(A \cap B) = \mathrm{Id}(A) \cup \mathrm{Id}(B)$, for the union of two ideals is almost never an ideal.

The next idea arises in characterizing those ideals of the form $\mathrm{Id}(V)$ when V is a variety.

Definition. If I is an ideal in a commutative ring R, then its ***radical***, denoted by $\sqrt{I}$, is

$$\sqrt{I} = \{r \in R : r^m \in I \text{ for some integer } m \geq 1\}.$$

An ideal I is called a ***radical ideal***[14] if

$$\sqrt{I} = I.$$

Exercise 6.62 on page 397 asks you to prove that $\sqrt{I}$ is an ideal. It is easy to see that $I \subseteq \sqrt{I}$, and so an ideal I is a radical ideal if and only if $\sqrt{I} \subseteq I$. For example, every prime ideal P is a radical ideal, for if $f^n \in P$, then $f \in P$. Here is an example of an ideal that is not radical. Let $b \in k$ and let $I = ((x - b)^2)$. Now I is not a radical ideal, for $(x - b)^2 \in I$ while $x - b \notin I$.

Definition. An element a in a commutative ring R is called ***nilpotent*** if $a \neq 0$ and there is some $n \geq 1$ with $a^n = 0$.

Note that I is a radical ideal in a commutative ring R if and only if R/I has no nonzero nilpotent elements. A commutative ring having no nilpotent elements is called ***reduced***.

[14]This term is appropriate, for if $r^m \in I$, then its mth root r also lies in I.

Proposition 6.97. *If an ideal $I = \mathrm{Id}(A)$ for some $A \subseteq k^n$, then it is a radical ideal. Hence, the coordinate ring $k[A]$ has no nonzero nilpotent elements.*

Proof. Since $I \subseteq \sqrt{I}$ is always true, it suffices to check the reverse inclusion. By hypothesis, $I = \mathrm{Id}(A)$ for some $A \subseteq k^n$; hence, if $f \in \sqrt{I}$, then $f^m \in \mathrm{Id}(A)$; that is, $f(a)^m = 0$ for all $a \in A$. But the values of $f(a)^m$ lie in the field k, and so $f(a)^m = 0$ implies $f(a) = 0$; that is, $f \in \mathrm{Id}(A) = I$. •

Proposition 6.98.

(i) *If I and J are ideals, then $\sqrt{I \cap J} = \sqrt{I} \cap \sqrt{J}$.*

(ii) *If I and J are radical ideals, then $I \cap J$ is a radical ideal.*

Proof. (i) If $f \in \sqrt{I \cap J}$, then $f^m \in I \cap J$ for some $m \geq 1$. Hence, $f^m \in I$ and $f^m \in J$, and so $f \in \sqrt{I}$ and $f \in \sqrt{J}$; that is, $f \in \sqrt{I} \cap \sqrt{J}$.

For the reverse inclusion, assume that $f \in \sqrt{I} \cap \sqrt{J}$, so that $f^m \in I$ and $f^q \in J$. We may assume that $m \geq q$, and so $f^m \in I \cap J$; that is, $f \in \sqrt{I \cap J}$.

(ii) If I and J are radical ideals, then $I = \sqrt{I}$ and $J = \sqrt{J}$ and

$$I \cap J \subseteq \sqrt{I \cap J} = \sqrt{I} \cap \sqrt{J} = I \cap J. \quad \bullet$$

We are now going to prove Hilbert's *Nullstellensatz* for $\mathbb{C}[X]$. The reader will see that the proof we will give generalizes to any uncountable algebraically closed field. The theorem is actually true for all algebraically closed fields (we shall prove it in Chapter 11), and so the proof here does not, alas, cover the algebraic closures of the prime fields, for example, which are countable.

Lemma 6.99. *Let k be a field and let $\varphi \colon k[X] \to k$ be a surjective ring homomorphism which fixes k pointwise. If $J = \ker\varphi$, then $\mathrm{Var}(J) \neq \varnothing$.*

Proof. Let $\varphi(x_i) = a_i \in k$ and let $a = (a_1, \ldots, a_n) \in k^n$. If

$$f(X) = \sum_{\alpha_1,\ldots,\alpha_n} c_{\alpha_1,\ldots,\alpha_n} x_1^{\alpha_1} \cdots x_n^{\alpha_n} \in k[X],$$

then

$$\begin{aligned}
\varphi(f(X)) &= \sum_{\alpha_1,\ldots,\alpha_n} c_{\alpha_1,\ldots,\alpha_n} \varphi(x_1)^{\alpha_1} \cdots \varphi(x_n)^{\alpha_n} \\
&= \sum_{\alpha_1,\ldots,\alpha_n} c_{\alpha_1,\ldots,\alpha_n} a_1^{\alpha_1} \cdots a_n^{\alpha_n} \\
&= f(a_1, \ldots, a_n) \\
&= f(a).
\end{aligned}$$

Hence, if $f(X) \in J = \ker\varphi$, then $f(a) = 0$, and so $a \in \mathrm{Var}(J)$. •

The next proof will use a bit of cardinality.

Theorem 6.100 (Weak Nullstellensatz[15] over $\mathbb{C}$). *If $f_1(X), \ldots, f_t(X) \in \mathbb{C}[X]$, then $I = (f_1, \ldots, f_t)$ is a proper ideal in $\mathbb{C}[X]$ if and only if* $\mathrm{Var}(f_1, \ldots, f_t) \neq \varnothing$.

Remark. The reader should note that the only properties of $\mathbb{C}$ used in the proof are that it is an uncountable algebraically closed field. ◂

Proof. It is clear that if $\mathrm{Var}(I) \neq \varnothing$, then I is a proper ideal, because $\mathrm{Var}(\mathbb{C}[X]) = \varnothing$.

For the converse, suppose that I is a proper ideal. By Corollary 6.40, there is a maximal ideal M containing I, and so $K = \mathbb{C}[X]/M$ is a field. It is plain that the natural map $\mathbb{C}[X] \to \mathbb{C}[X]/M = K$ carries $\mathbb{C}$ to itself, so that $K/\mathbb{C}$ is an extension field; it follows that K is a vector space over $\mathbb{C}$. Now $\mathbb{C}[X]$ has countable dimension, as a $\mathbb{C}$-space, for a basis consists of all the monic monomials $1, x, x^2, x^3, \ldots$. Therefore, $\dim_{\mathbb{C}}(K)$ is countable (possibly finite), for it is a quotient of $\mathbb{C}[X]$.

Suppose that K is a proper extension of $\mathbb{C}$; that is, there is some $t \in K$ with $t \notin \mathbb{C}$. Since $\mathbb{C}$ is algebraically closed, t cannot be algebraic over $\mathbb{C}$, and so it is transcendental. Consider the subset B of K,

$$B = \{1/(t-c) \colon c \in \mathbb{C}\}$$

(note that $t - c \neq 0$ because $t \notin \mathbb{C}$). The set B is uncountable, for it is indexed by the uncountable set $\mathbb{C}$. We claim that B is linearly independent over $\mathbb{C}$; if so, then the fact that $\dim_{\mathbb{C}}(K)$ is countable is contradicted, and we will conclude that $K = \mathbb{C}$. If B is linearly dependent, there are nonzero $a_1, \ldots, a_r \in \mathbb{C}$ and distinct $c_1, \ldots, c_r \in \mathbb{C}$ with $\sum_{i=1}^{r} a_i/(t - c_i) = 0$. Clearing denominators, we have a polynomial $h(t) \in \mathbb{C}[t]$:

$$h(t) = \sum_i a_i (t - c_1) \cdots \widehat{(t - c_i)} \cdots (t - c_r) = 0.$$

Now $h(c_1) = a_1(c_1 - c_2) \cdots (c_1 - c_r) \neq 0$, so that $h(t)$ is not the zero polynomial. But this contradicts t being transcendental; therefore, $K = \mathbb{C}$. Lemma 6.99 now applies to show that $\mathrm{Var}(M) \neq \varnothing$. But $\mathrm{Var}(M) \subseteq \mathrm{Var}(I)$, and this completes the proof. •

Consider the special case of this theorem for $I = (f(x)) \subseteq \mathbb{C}[x]$, where $f(x)$ is not a constant. To say that $\mathrm{Var}(f) \subseteq \mathbb{C}$ is nonempty is to say that $f(x)$ has a complex root. Thus, the weak Nullstellensatz is a generalization to several variables of the fundamental theorem of algebra.

Theorem 6.101. *If k is an* (*uncountable*) *algebraically closed field, then every maximal ideal M in $k[x_1, \ldots, x_n]$ has the form*

$$M = (x_1 - a_1, \ldots, x_n - a_n),$$

where $a = (a_1, \ldots, a_n) \in k^n$, and so there is a bijection between k^n and the maximal ideals in $k[x_1, \ldots, x_n]$.

[15]The German word *Nullstelle* means *root*. In the context of polynomials of several variables, we may translate it as *zero*, and so *Nullstellensatz* means the *theorem of zeros*.

Remark. The uncountability hypothesis will be removed in Chapter 11. ◀

Proof. Since $k[X]/M \cong k$, Lemma 6.99 gives $\mathrm{Var}(M) \neq \varnothing$. As in the proof of that lemma, there are constants $a_i \in k$ with $x_i + M = a_i + M$ for all i, and so $x_i - a_i \in M$. Therefore, there is an inclusion of ideals

$$(x_1 - a_1, \ldots, x_n - a_n) \subseteq M.$$

But $(x_1 - a_1, \ldots, x_n - a_n)$ is a maximal ideal, by Exercise 6.6(i) on page 325, and so $M = (x_1 - a_1, \ldots, x_n - a_n)$. •

The following proof of Hilbert's Nullstellensatz uses the "Rabinowitch trick" of imbedding a polynomial ring in n variables into a polynomial ring in $n+1$ variables. Again, uncountability is not needed, and we assume it only because our proof of the weak Nullstellensatz uses this hypothesis.

Theorem 6.102 (Nullstellensatz). *Let k be an* (*uncountable*) *algebraically closed field. If I is an ideal in $k[X]$, then* $\mathrm{Id}(\mathrm{Var}(I)) = \sqrt{I}$. *Thus, f vanishes on* $\mathrm{Var}(I)$ *if and only if $f^m \in I$ for some $m \geq 1$.*

Proof. The inclusion $\mathrm{Id}(\mathrm{Var}(I)) \supseteq \sqrt{I}$ is obviously true, for if $f^m(a) = 0$ for some $m \geq 1$ and all $a \in \mathrm{Var}(I)$, then $f(a) = 0$ for all a, because $f(a) \in k$.

For the converse, assume that $h \in \mathrm{Id}(\mathrm{Var}(I))$, where $I = (f_1, \ldots, f_t)$; that is, if $f_i(a) = 0$ for all i, where $a \in k^n$, then $h(a) = 0$. We must show that some power of h lies in I. Of course, we may assume that h is not the zero polynomial. Let us regard

$$k[x_1, \ldots, x_n] \subseteq k[x_1, \ldots, x_n, y];$$

thus, every $f_i(x_1, \ldots, x_n)$ is regarded as a polynomial in $n+1$ variables that does not depend on the last variable y. We claim that the polynomials

$$f_1, \ldots, f_t, 1 - yh$$

in $k[x_1, \ldots, x_n, y]$ have no common zeros. If $(a_1, \ldots, a_n, b) \in k^{n+1}$ is a common zero, then $a = (a_1, \ldots, a_n) \in k^n$ is a common zero of $f_1, \ldots, f_t$, and so $h(a) = 0$. But now $1 - bh(a) = 1 \neq 0$. The weak Nullstellensatz now applies to show that the ideal $(f_1, \ldots, f_t, 1 - yh)$ in $k[x_1, \ldots, x_n, y]$ is not a proper ideal. Therefore, there are $g_1, \ldots,$ $g_{t+1} \in k[x_1, \ldots, x_n, y]$ with

$$1 = f_1 g_1 + \cdots + f_t g_t + (1 - yh) g_{t+1}.$$

Make the substitution $y = 1/h$, so that the last term involving g_{t+1} vanishes. Rewriting, $g_i(X, y) = \sum_{j=0}^{d_i} u_j(X) y^j$, and so $g_i(X, h^{-1}) = \sum_{j=0}^{d_i} u_j(X) h^{-j}$. It follows that

$$h^{d_i} g_i(X, h^{-1}) \in k[X].$$

Therefore, if $m = \max\{d_1, \ldots, d_t\}$, then

$$h^m = (h^m g_1) f_1 + \cdots + (h^m g_t) f_t \in I. \quad \bullet$$

We continue the study of the operators Var and Id.

Proposition 6.103. *Let k be any field.*

(i) *For every subset $F \subseteq k^n$,*
$$\mathrm{Var}(\mathrm{Id}(F)) \supseteq F.$$

(ii) *For every ideal $I \subseteq k[X]$,*
$$\mathrm{Id}(\mathrm{Var}(I)) \supseteq I.$$

(iii) *If V is a variety of k^n, then* $\mathrm{Var}(\mathrm{Id}(V)) = V$.

(iv) *If F is subset of k^n, then $\overline{F}$, the intersection of all those varieties that contain F, is equal to* $\mathrm{Var}(\mathrm{Id}(F))$. *One calls $\overline{F}$ the* ***Zariski closure***[16] *of F.*

(v) *If $V \subseteq V^* \subseteq k^n$ are varieties, then*
$$V^* = V \cup \overline{V^* - V},$$
the Zariski closure of $V^* - V$.

Proof. (i) This result is almost a tautology. If $a \in F$, then $g(a) = 0$ for all $g(X) \in \mathrm{Id}(F)$. But every $g(X) \in \mathrm{Id}(F)$ annihilates F, by definition of $\mathrm{Id}(F)$, and so $a \in \mathrm{Var}(\mathrm{Id}(F))$. Therefore, $\mathrm{Var}(\mathrm{Id}(F)) \supseteq F$.

(ii) Again, we merely look at the definitions. If $f(X) \in I$, then $f(a) = 0$ for all $a \in \mathrm{Var}(I)$; hence, $f(X)$ is surely one of the polynomials annihilating $\mathrm{Var}(I)$.

(iii) If V is a variety, then $V = \mathrm{Var}(J)$ for some ideal J in $k[X]$. Now
$$\mathrm{Var}(\mathrm{Id}(\mathrm{Var}(J))) \supseteq \mathrm{Var}(J),$$
by part (i). Also, part (ii) gives $\mathrm{Id}(\mathrm{Var}(J)) \supseteq J$, and applying Proposition 6.92(i) gives the reverse inclusion
$$\mathrm{Var}(\mathrm{Id}(\mathrm{Var}(J))) \subseteq \mathrm{Var}(J).$$
Therefore, $\mathrm{Var}(\mathrm{Id}(\mathrm{Var}(J))) = \mathrm{Var}(J)$; that is, $\mathrm{Var}(\mathrm{Id}(V)) = V$.

(iv) By Proposition 6.93(iii), $\overline{F} = \bigcap_{V \supseteq F} V$ is a variety containing F. Since $\mathrm{Var}(\mathrm{Id}(F))$ is a variety containing F, it is one of varieties V being intersected to form $\overline{F}$, and so $\overline{F} \subseteq \mathrm{Var}(\mathrm{Id}(F))$. For the reverse inclusion, it suffices to prove that if V is any variety containing F, then $V \supseteq \mathrm{Var}(\mathrm{Id}(F))$. If $V \supseteq F$, then $\mathrm{Id}(V) \subseteq \mathrm{Id}(F)$, and $V = \mathrm{Var}(\mathrm{Id}(V)) \supseteq \mathrm{Var}(\mathrm{Id}(F))$.

(v) Since $V^* - V \subseteq V^*$, we have $\overline{V^* - V} \subseteq \overline{V^*} = V^*$. By hypothesis, $V \subseteq V^*$, and so $V \cup \overline{V^* - V} \subseteq V^*$. For the reverse inclusion, there is an equation of subsets, $V^* = V \cup (V^* - V)$. Taking closures,
$$V^* = \overline{V^*} = \overline{V} \cup \overline{V^* - V} = V \cup \overline{V^* - V},$$
because $V = \overline{V}$. •

[16] If F is a subset of a topological space X, then its ***closure*** is defined as the intersection of all the closed sets in X that contain F.

Corollary 6.104.

(i) *If* V_1 *and* V_2 *are varieties and* $\mathrm{Id}(V_1) = \mathrm{Id}(V_2)$, *then* $V_1 = V_2$.

(ii) *Let* k *be an* (*uncountable*) *algebraically closed field. If* I_1 *and* I_2 *are radical ideals and* $\mathrm{Var}(I_1) = \mathrm{Var}(I_2)$, *then* $I_1 = I_2$.

Proof. (i) If $\mathrm{Id}(V_1) = \mathrm{Id}(V_2)$, then $\mathrm{Var}(\mathrm{Id}(V_1)) = \mathrm{Var}(\mathrm{Id}(V_2))$; it now follows from Proposition 6.103(iii) that $V_1 = V_2$.

(ii) If $\mathrm{Var}(I_1) = \mathrm{Var}(I_2)$, then $\mathrm{Id}(\mathrm{Var}(I_1)) = \mathrm{Id}(\mathrm{Var}(I_2))$. By the Nullstellensatz, which holds because k is an (uncountable) algebraically closed field, $\sqrt{I_1} = \sqrt{I_2}$. Since I_1 and I_2 are radical ideals, by hypothesis, we have $I_1 = I_2$. •

Can a variety be decomposed into simpler subvarieties?

Definition. A variety V is ***irreducible*** if it is not a union of two proper subvarieties; that is, $V \neq W' \cup W''$, where both W' and W'' are varieties that are proper subsets of V.

Proposition 6.105. *Every variety* V *in* k^n *is a union of finitely many irreducible subvarieties:*

$$V = V_1 \cup V_2 \cup \cdots \cup V_m.$$

Proof. Call a variety $W \in k^n$ *good* if it is irreducible or a union of finitely many irreducible subvarieties; otherwise, call W *bad*. We must show that there are no bad varieties. If W is bad, it is not irreducible, and so $W = W' \cup W''$, where both W' and W'' are proper subvarieties. But a union of good varieties is good, and so at least one of W' and W'' is bad; say, W' is bad, and rename it $W' = W_1$. Repeat this construction for W_1 to get a bad subvariety W_2. It follows by induction that there exists a strictly descending sequence

$$W \supsetneq W_1 \supsetneq \cdots \supsetneq W_n \supsetneq \cdots$$

of bad subvarieties. Since the operator Id reverses inclusions, there is a strictly increasing chain of ideals

$$\mathrm{Id}(W) \subsetneq \mathrm{Id}(W_1) \subsetneq \cdots \subsetneq \mathrm{Id}(W_n) \subsetneq \cdots$$

[the inclusions are strict because of Corollary 6.104(i)], and this contradicts the Hilbert basis theorem. We conclude that every variety is good. •

Irreducible varieties have a nice characterization.

Proposition 6.106. *A variety* V *in* k^n *is irreducible if and only if* $\mathrm{Id}(V)$ *is a prime ideal in* $k[X]$. *Hence, the coordinate ring* $k[V]$ *of an irreducible variety* V *is a domain.*

Proof. Assume that V is an irreducible variety. It suffices to show that if $f_1(X), f_2(X) \notin \mathrm{Id}(V)$, then $f_1(X) f_2(X) \notin \mathrm{Id}(V)$. Define, for $i = 1, 2$,

$$W_i = V \cap \mathrm{Var}(f_i(X)).$$

Note that each W_i is a subvariety of V, for it is the intersection of two varieties; moreover, since $f_i(X) \notin \mathrm{Id}(V)$, there is some $a_i \in V$ with $f_i(a_i) \neq 0$, and so W_i is a proper subvariety of V. Since V is irreducible, we cannot have $V = W_1 \cup W_2$. Thus, there is some $b \in V$ that is not in $W_1 \cup W_2$; that is, $f_1(b) \neq 0 \neq f_2(b)$. Therefore, $f_1(b) f_2(b) \neq 0$, hence $f_1(X) f_2(X) \notin \mathrm{Id}(V)$, and so $\mathrm{Id}(V)$ is a prime ideal.

Conversely, assume that $\mathrm{Id}(V)$ is a prime ideal. Suppose that $V = V_1 \cup V_2$, where V_1 and V_2 are subvarieties. If $V_2 \subsetneq V$, then we must show that $V = V_1$. Now

$$\mathrm{Id}(V) = \mathrm{Id}(V_1) \cap \mathrm{Id}(V_2) \supseteq \mathrm{Id}(V_1)\,\mathrm{Id}(V_2);$$

the equality is given by Proposition 6.96, and the inequality is given by Exercise 6.10 on page 325. Since $\mathrm{Id}(V)$ is a prime ideal, Proposition 6.13 says that $\mathrm{Id}(V_1) \subseteq \mathrm{Id}(V)$ or $\mathrm{Id}(V_2) \subseteq \mathrm{Id}(V)$. But $V_2 \subsetneq V$ implies $\mathrm{Id}(V_2) \supsetneq \mathrm{Id}(V)$, and we conclude that $\mathrm{Id}(V_1) \subseteq \mathrm{Id}(V)$. Now the reverse inequality $\mathrm{Id}(V_1) \supseteq \mathrm{Id}(V)$ holds as well, because $V_1 \subseteq V$, and so $\mathrm{Id}(V_1) = \mathrm{Id}(V)$. Therefore, $V_1 = V$, by Corollary 6.104, and so V is irreducible. •

We now consider whether the irreducible subvarieties in the decomposition of a variety into a union of irreducible varieties are uniquely determined. There is one obvious way to arrange nonuniqueness. If $P \subsetneq Q$ in $k[X]$ are two prime ideals (for example, $(x) \subsetneq (x, y)$ are such prime ideals in $k[x, y]$), then $\mathrm{Var}(Q) \subsetneq \mathrm{Var}(P)$; if $\mathrm{Var}(P)$ is a subvariety of a variety V, say, $V = \mathrm{Var}(P) \cup V_2 \cup \cdots \cup V_m$, then $\mathrm{Var}(Q)$ can be one of the V_i or it can be left out.

Definition. A decomposition $V = V_1 \cup \cdots \cup V_m$ is an ***irredundant union*** if no V_i can be omitted; that is, for all i,

$$V \neq V_1 \cup \cdots \cup \widehat{V_i} \cup \cdots \cup V_m.$$

Proposition 6.107. *Every variety V is an irredundant union of irreducible subvarieties*

$$V = V_1 \cup \cdots \cup V_m;$$

moreover, the irreducible subvarieties V_i are uniquely determined by V.

Proof. By Proposition 6.105, V is a union of finitely many irreducible subvarieties; say, $V = V_1 \cup \cdots \cup V_m$. If m is chosen minimal, then this union must be irredundant.

We now prove uniqueness. Suppose that $V = W_1 \cup \cdots \cup W_s$ is an irredundant union of irreducible subvarieties. Let $X = \{V_1, \ldots, V_m\}$ and let $Y = \{W_1, \ldots, W_s\}$; we shall show that $X = Y$. If $V_i \in X$, we have

$$V_i = V_i \cap V = \bigcup_j (V_i \cap W_j).$$

Now $V_i \cap W_j \neq \varnothing$ for some j; since V_i is irreducible, there is only one such W_j. Therefore, $V_i = V_i \cap W_j$, and so $V_i \subseteq W_j$. The same argument applied to W_j shows that there is exactly one V_ℓ with $W_j \subseteq V_\ell$. Hence,

$$V_i \subseteq W_j \subseteq V_\ell.$$

Since the union $V_1 \cup \cdots \cup V_m$ is irredundant, we must have $V_i = V_\ell$, and so $V_i = W_j = V_\ell$; that is, $V_i \in Y$ and $X \subseteq Y$. The reverse inclusion is proved in the same way. •

Definition. An intersection $I = J_1 \cap \cdots \cap J_m$ is ***irredundant*** if no J_i can be omitted; that is, for all i,

$$I \neq J_1 \cap \cdots \cap \widehat{J_i} \cap \cdots \cap J_m.$$

Corollary 6.108. *Every radical ideal J in $k[X]$ is an irredundant intersection of prime ideals,*

$$J = P_1 \cap \cdots \cap P_m;$$

moreover, the prime ideals P_i are uniquely determined by J.

Remark. This corollary is generalized in Exercise 6.72 on page 399: An ideal in an arbitrary commutative noetherian ring is a radical ideal if and only if it is an intersection of finitely many prime ideals. ◂

Proof. Since J is a radical ideal, there is a variety V with $J = \mathrm{Id}(V)$. Now V is an irredundant union of irreducible subvarieties,

$$V = V_1 \cup \cdots \cup V_m,$$

so that

$$J = \mathrm{Id}(V) = \mathrm{Id}(V_1) \cap \cdots \cap \mathrm{Id}(V_m).$$

By Proposition 6.106, V_i irreducible implies $\mathrm{Id}(V_i)$ is prime, and so J is an intersection of prime ideals. This is an irredundant intersection, for if there is ℓ with $J = \mathrm{Id}(V) = \bigcap_{j \neq \ell} \mathrm{Id}(V_j)$, then

$$V = \mathrm{Var}(\mathrm{Id}(V)) = \bigcup_{j \neq \ell} \mathrm{Var}(\mathrm{Id}(V_j)) = \bigcup_{j \neq \ell} V_j,$$

contradicting the given irredundancy of the union.

Uniqueness is proved similarly. If $J = \mathrm{Id}(W_1) \cap \cdots \cap \mathrm{Id}(W_s)$, where each $\mathrm{Id}(W_i)$ is a prime ideal (hence is a radical ideal), then each W_i is an irreducible variety. Applying Var expresses $V = \mathrm{Var}(\mathrm{Id}(V)) = \mathrm{Var}(J)$ as an irredundant union of irreducible subvarieties, and the uniqueness of this decomposition gives the uniqueness of the prime ideals in the intersection. •

Given an ideal I in $k[x_1, \ldots, x_n]$, how can we find the irreducible components C_i of $\mathrm{Var}(I)$? To ask the question another way, what are the prime ideals P_i with $C_i = \mathrm{Var}(P_i)$? The first guess is that $I = P_1 \cap \cdots \cap P_r$, but this is easily seen to be incorrect: There are ideals I that are not an intersection of prime ideals. For example, in $k[x]$, the ideal $((x-1)^2)$ is not an intersection of prime ideals. In light of the Nullstellensatz, we can replace the prime ideals P_i by ideals Q_i with $\sqrt{Q_i} = P_i$, for $\mathrm{Var}(P_i) = \mathrm{Var}(Q_i)$. We are

led to the notion of *primary ideal*, defined soon, and the *primary decomposition theorem*, which states that every ideal in a commutative noetherian ring, not merely in $k[X]$, is an intersection of primary ideals.

We can now give a geometric interpretation of the colon ideal.

Proposition 6.109. *Let k be an (uncountable) algebraically closed field, and let I be a radical ideal in $k[X]$. Then, for every ideal J,*

$$\mathrm{Var}((I : J)) = \overline{\mathrm{Var}(I) - \mathrm{Var}(J)}.$$

Proof. We first show that $\mathrm{Var}((I : J)) \supseteq \overline{\mathrm{Var}(I) - \mathrm{Var}(J)}$. If $f \in (I : J)$, then $fg \in I$ for all $g \in J$. Hence, if $x \in \mathrm{Var}(I)$, then $f(x)g(x) = 0$ for all $g \in J$. However, if $x \notin \mathrm{Var}(J)$, then there is $g \in J$ with $g(x) \neq 0$. Since $k[X]$ is a domain, we have $f(x) = 0$ for all $x \in \mathrm{Var}(I) - \mathrm{Var}(J)$; that is, $f \in \mathrm{Id}(\mathrm{Var}(I) - \mathrm{Var}(J))$. Thus, $(I : J) \subseteq \mathrm{Id}(\mathrm{Var}(I) - \mathrm{Var}(J))$, and so

$$\mathrm{Var}((I : J)) \supseteq \mathrm{Var}(\mathrm{Id}(\mathrm{Var}(I) - \mathrm{Var}(J))) = \overline{\mathrm{Var}(I) - \mathrm{Var}(J)},$$

by Proposition 6.103(iv).

For the reverse inclusion, take $x \in \mathrm{Var}((I : J))$. Thus, if $f \in (I : J)$, then $f(x) = 0$; that is,

$$\text{if } fg \in I \text{ for all } g \in J, \text{ then } f(x) = 0.$$

Suppose now that $h \in \mathrm{Id}(\mathrm{Var}(I) - \mathrm{Var}(J))$. If $g \in J$, then hg vanishes on $\mathrm{Var}(J)$ (because g does); on the other hand, hg vanishes on $\mathrm{Var}(I) - \mathrm{Var}(J)$ (because h does). It follows that hg vanishes on $\mathrm{Var}(J) \cup (\mathrm{Var}(I) - \mathrm{Var}(J)) = \mathrm{Var}(I)$; hence, $hg \in \sqrt{I} = I$ for all $g \in J$, because I is a radical ideal, and so $h \in (I : J)$. Therefore, $h(x) = 0$ for all $h \in (I : J)$, which gives $x \in \mathrm{Var}(\mathrm{Id}(\mathrm{Var}(I) - \mathrm{Var}(J))) = \overline{\mathrm{Var}(I) - \mathrm{Var}(J)}$, as desired. •

Definition. An ideal Q in a commutative ring R is ***primary*** if it is a proper ideal and if $ab \in Q$ (where $a, b \in R$) and $b \notin Q$, then $a^n \in Q$ for some $n \geq 1$.

It is clear that every prime ideal is primary. Moreover, in $\mathbb{Z}$, the ideal (p^e), where p is prime and $e \geq 2$, is a primary ideal that is not a prime ideal. Example 6.114 shows that this example is misleading: There are primary ideals that are not powers of prime ideals; there are powers of prime ideals which are not primary ideals.

Proposition 6.110. *If Q is a primary ideal, then its radical $P = \sqrt{Q}$ is a prime ideal. Moreover, if Q is primary, then $ab \in Q$ and $a \notin Q$ implies $b \in P$.*

Proof. Assume that $ab \in \sqrt{Q}$, so that $(ab)^m = a^m b^m \in Q$ for some $m \geq 1$. If $a \notin \sqrt{Q}$, then $a^m \notin Q$. Since Q is primary, it follows that some power of b^m, say, $b^{mn} \in Q$; that is, $b \in \sqrt{Q}$. We have proved that $\sqrt{Q}$ is prime, as well as the second statement. •

If Q is primary and $P = \sqrt{Q}$, then we often call Q a ***P-primary ideal***, and we say that Q and P ***belong*** to each other.

We now prove that the properties in Proposition 6.110 characterize primary ideals.

Proposition 6.111. *Let J and T be ideals in a commutative ring. If* (i) $J \subseteq T$, (ii) $t \in T$ *implies there is some* $m \geq 1$ *with* $t^m \in J$, *and* (iii) *if* $ab \in J$ *and* $a \notin J$, *then* $b \in T$, *then J is a primary ideal with radical T.*

Proof. First, J is a primary ideal, for if $ab \in J$ and $a \notin J$, then axiom (iii) gives $b \in T$, and axiom (ii) gives $b^m \in J$. It remains to prove that $T = \sqrt{J}$. Now axiom (ii) gives $T \subseteq \sqrt{J}$. For the reverse inclusion, if $r \in \sqrt{J}$, then $r^m \in J$; choose m minimal. If $m = 1$, then axiom (i) gives $r \in J \subseteq T$, as desired. If $m > 1$, then $rr^{m-1} \in J$; since $r^{m-1} \notin J$, axiom (iii) gives $r \in T$. Therefore, $T = \sqrt{J}$. •

Let R be a commutative ring, and let M be an ideal. Each $a \in R$ defines an R-map $a_M : M \to M$ by $a_M : m \mapsto am$.

Lemma 6.112. *Let Q be an ideal in a commutative ring R. Then Q is a primary ideal if and only if, for each $a \in R$, the map $a_{R/Q} : R/Q \to R/Q$, given by $r + Q \mapsto ar + Q$, is either an injection or is nilpotent* $[(a_{R/Q})^n = 0$ *for some* $n \geq 1]$.

Proof. Assume that Q is primary. If $a \in R$ and $a_{R/Q}$ is not an injection, then there is $b \in R$ with $b \notin Q$ and $a_{R/Q}(b + Q) = ab + Q = Q$; that is, $ab \in Q$. We must prove that $a_{R/Q}$ is nilpotent. Since Q is primary, there is $n \geq 1$ with $a^n \in Q$; hence, $a^n r \in Q$ for all $r \in R$, because Q is an ideal. Thus, $(a_{R/Q})^n (r + Q) = a^n r + Q = Q$ for all $r \in R$, and $(a_{R/Q})^n = 0$; that is, $a_{R/Q}$ is nilpotent.

Conversely, assume that every $a_{R/Q}$ is either injective or nilpotent. Suppose that $ab \in Q$ and $a \notin Q$. Then $b_{R/Q}$ is not injective, for $a + Q \in \ker b_{R/Q}$. By hypothesis, $(b_{R/Q})^n = 0$ for some $n \geq 1$; that is, $b^n r \in Q$ for all $r \in R$. Setting $r = 1$ gives $b^n \in Q$, and so Q is primary. •

The next result gives a way of constructing primary ideals.

Proposition 6.113. *If P is a maximal ideal in a commutative ring R, and if Q is an ideal with $P^e \subseteq Q \subseteq P$ for some $e \geq 0$, then Q is a P-primary ideal. In particular, every power of a maximal ideal is primary.*

Proof. We show, for each $a \in R$, that $a_{R/Q}$ is either nilpotent or injective. Suppose first that $a \in P$. In this case, $a^e \in P^e \subseteq Q$; hence, $a^e b \in Q$ for all $b \in R$, and so $(a_{R/Q})^e = 0$; that is, $a_{R/Q}$ is nilpotent. Now assume that $a \notin P$; we are going to show that $a + Q$ is a unit in R/Q, which implies that $a_{R/Q}$ is injective. Since P is a maximal ideal, the ring R/P is a field; since $a \notin P$, the element $a + P$ is a unit in R/P: there is $a' \in R$ and $z \in P$ with $aa' = 1 - z$. Now $z + Q$ is a nilpotent element of R/Q, for $z^e \in P^e \subseteq Q$. Thus, $1 - z + Q$ is a unit in R/Q (its inverse is $1 + z + \cdots + z^{e-1}$). It follows that $a + Q$ is a unit in R/Q, for $aa' + Q = 1 - z + Q$. The result now follows from Lemma 6.112. Finally, Q belongs to P, for $P = \sqrt{P^e} \subseteq \sqrt{Q} \subseteq \sqrt{P} = P$. •

Example 6.114.

(i) We now show that a power of a prime ideal need not be primary. Suppose that R is a commutative ring containing elements a, b, c such that $ab = c^2$, $P = (a, c)$ is a prime

ideal, $a \notin P^2$, and $b \notin P$. Now $ab = c^2 \in P^2$; were P^2 primary, then $a \notin P^2$ would imply that $b \in \sqrt{P^2} = P$, and this is not so. We construct such a ring R as follows. Let k be a field, and define $R = k[x, y, z]/(xy - z^2)$ (note that R is noetherian). Define $a, b, c \in R$ to be the cosets of x, y, z, respectively. Now $P = (a, c)$ is a prime ideal, for the third isomorphism theorem for rings, Exercise 3.82 on page 196, gives

$$R/(a, c) = \frac{k[x, y, z]/(xy - z^2)}{(x, z)/(xy - z^2)} \cong \frac{k[x, y, z]}{(x, z)} \cong k[y],$$

which is a domain. The equation $ab = c^2$ obviously holds in R. Were $a \in P^2$, then lifting this relation to $k[x, y, z]$ would yield an equation

$$x = f(x, y, z)x^2 + g(x, y, z)xz + h(x, y, z)z^2 + \ell(x, y, z)(xy - z^2).$$

Setting $y = 0 = z$ (i.e., using the evaluation homomorphism $k[x, y, z] \to k[x]$) gives the equation $x = f(x, 0, 0)x^2$ in $k[x]$, a contradiction. A similar argument shows that $b \notin P$.

(ii) We use Proposition 6.113 to show that there are primary ideals Q that are not powers of prime ideals. Let $R = k[x, y]$, where k is a field. The ideal $P = (x, y)$ is maximal, hence prime (for $R/P \cong k$); moreover,

$$P^2 \subsetneq (x^2, y) \subsetneq (x, y) = P$$

[the strict inequalities follow from $x \notin (x^2, y)$ and $y \notin P^2$]. Thus, $Q = (x^2, y)$ is not a power of P; indeed, we show that $Q \neq L^e$, where L is a prime ideal. If $Q = L^e$, then $P^2 \subseteq L^e \subseteq P$, hence $\sqrt{P^2} \subseteq \sqrt{L^e} \subseteq \sqrt{P}$, and so $P \subseteq L \subseteq P$, a contradiction. ◂

We now generalize Corollary 6.108 by proving that every ideal in a noetherian ring, in particular, in $k[X]$ for k a field, is an intersection of primary ideals. This result, along with uniqueness properties, was first proved by E. Lasker; his proof was later simplified by E. Noether. Note that we will be working in arbitrary noetherian rings, not merely in $k[X]$.

Definition. A ***primary decomposition*** of an ideal I in a commutative ring R is a finite family of primary ideals $Q_1, \ldots, Q_r$ with

$$I = Q_1 \cap Q_2 \cap \cdots \cap Q_r.$$

Theorem 6.115 (Lasker–Noether I). *If R is a commutative noetherian ring, then every proper ideal I in R has a primary decomposition.*

Proof. Let $\mathcal{F}$ be the family of all those proper ideals in R that do not have a primary decomposition; we must show that $\mathcal{F}$ is empty. Since R is noetherian, if $\mathcal{F} \neq \varnothing$, then it has a maximal element, say, J. Of course, J is not primary, and so there exists $a \in R$ with $a_{R/J} \colon R/J \to R/J$ neither injective nor nilpotent. The ascending chain of ideals of R/J,

$$\ker\ a_{R/J} \subseteq \ker\ (a_{R/J})^2 \subseteq \ker\ (a_{R/J})^3 \subseteq \cdots,$$

must stop (because R/J, being a quotient of the noetherian ring R, is itself noetherian); there is $m \geq 1$ with $\ker(a_{R/J}^{\ell}) = \ker(a_{R/J}^{m})$ for all $\ell \geq m$. Denote $(a_{R/J})^m$ by φ, so that $\ker(\varphi^2) = \ker\varphi$. Note that $\ker\varphi \neq \{0\}$, because $\{0\} \subsetneq \ker a_{R/J} \subseteq \ker(a_{R/J})^m = \ker\varphi$, and that $\operatorname{im}\varphi = \operatorname{im}(a_{R/J})^m \neq \{0\}$, because $a_{R/J}$ is not nilpotent. We claim that

$$\ker\varphi \cap \operatorname{im}\varphi = \{0\}.$$

If $x \in \ker\varphi \cap \operatorname{im}\varphi$, then $\varphi(x) = 0$ and $x = \varphi(y)$ for some $y \in R/J$. But $\varphi(x) = \varphi(\varphi(y)) = \varphi^2(y)$, so that $y \in \ker(\varphi^2) = \ker\varphi$ and $x = \varphi(y) = 0$.

If $\pi\colon R \to R/J$ is the natural map, then $A = \pi^{-1}(\ker\varphi)$ and $A' = \pi^{-1}(\operatorname{im}\varphi)$ are ideals of R with $A \cap A' = J$. It is obvious that A is a proper ideal; we claim that A' is also proper. Otherwise, $A' = R$, so that $A \cap A' = A$; but $A \cap A' = J$, as we saw above, and $A \neq J$, a contradiction. Since A and A' are strictly larger than J, neither of them lies in $\mathcal{F}$: There are primary decompositions $A = Q_1 \cap \cdots \cap Q_m$ and $A' = Q'_1 \cap \cdots \cap Q'_n$. Therefore,

$$J = A \cap A' = Q_1 \cap \cdots \cap Q_m \cap Q'_1 \cap \cdots \cap Q'_n,$$

contradicting J not having a primary decomposition (for $J \in \mathcal{F}$). •

Definition. A primary decomposition $I = Q_1 \cap \cdots \cap Q_r$ is ***irredundant*** if no Q_i can be omitted; for all i,

$$I \neq Q_1 \cap \cdots \cap \widehat{Q_i} \cap \cdots \cap Q_r.$$

The prime ideals $P_1 = \sqrt{Q_1}, \ldots, P_r = \sqrt{Q_r}$ are called the ***associated prime ideals*** of the irredundant primary decomposition.

It is clear that any primary decomposition can be made irredundant by throwing away, one at a time, any primary ideals that contain the intersection of the others.

Theorem 6.116 (Lasker–Noether II). *If I is an ideal in a noetherian ring R, then any two irredundant primary decompositions of I have the same set of associated prime ideals. Hence, the associated prime ideals are uniquely determined by I.*

Proof. Let $I = Q_1 \cap \cdots \cap Q_r$ be an irredundant primary decomposition, and let $P_i = \sqrt{Q_i}$. We are going to prove that a prime ideal P in R is equal to some P_i if and only if there is $c \notin I$ with $(I : c)$ a P-primary ideal; this will suffice, for the colon ideal $(I : c)$ is defined solely in terms of I and not in terms of any primary decomposition.

Given P_i, there exists $c \in \bigcap_{j \neq i} Q_j$ with $c \notin Q_i$, because of irredundancy; we show that $(I : c_i)$ is P_i-primary. Recall Proposition 6.111: If the following three conditions hold: (i) $(I : c) \subseteq P_i$; (ii) $b \in P_i$ implies there is some $m \geq 1$ with $b^m \in (I : c)$; and (iii) if $ab \in (I : c)$ and $a \notin (I : c)$, then $b \in P_i$ and $(I : c)$ is P_i-primary.

To see (i), if $u \in (I : c)$, then $uc \in I \subseteq P_i$. As $c \notin Q_i$, we have $u \in P_i$, by Proposition 6.110. To prove (ii), we first show that $Q_i \subseteq (I : c)$. If $a \in Q_i$, then $ca \in Q_i$, since Q_i is an ideal. If $j \neq i$, then $c \in Q_j$, and so $ca \in Q_j$. Therefore,

$ca \in Q_1 \cap \cdots \cap Q_r = I$, and so $a \in (I : c)$. If, now, $b \in P_i$, then $b^m \in Q_i \subseteq (I : c)$. Finally, we establish (iii) by proving its contrapositive: If $xy \in (I : c)$ and $x \notin P_i$, then $y \in (I : c)$. Thus, assume that $xyc \in I$; since $I \subseteq Q_i$ and $x \notin P_i = \sqrt{Q_i}$, we have $yc \in Q_i$. But $yc \in Q_j$ for all $j \neq i$, for $c \in Q_j$. Therefore, $yc \in Q_1 \cap \cdots \cap Q_r = I$, and so $y \in (I : c)$. We conclude that $(I : c)$ is P_i-primary.

Conversely, assume that there is an element $c \notin I$ and a prime ideal P such that $(I : c)$ is P-primary. We must show that $P = P_i$ for some i. Exercise 6.14(ii) on page 326 gives $(I : c) = (Q_1 : c) \cap \cdots \cap (Q_r : c)$. Therefore, by Proposition 6.98,

$$P = \sqrt{(I : c)} = \sqrt{(Q_1 : c)} \cap \cdots \cap \sqrt{(Q_r : c)}.$$

If $c \in Q_i$, then $(Q_i : c) = R$; if $c \notin Q_i$, then we saw, in first part of this proof, that $(Q_i : c)$ is P_i-primary. Thus, there is $s \leq r$ with

$$P = \sqrt{(Q_{i_1} : c)} \cap \cdots \cap \sqrt{(Q_{i_s} : c)} = P_{i_1} \cap \cdots \cap P_{i_s}.$$

Of course, $P \subseteq P_{i_j}$ for all j. On the other hand, Exercise 6.10(iii) on page 325 gives $P_{i_j} \subseteq P$ for some j, and so $P = P_{i_j}$, as desired. •

Example 6.117.
(i) Let $R = \mathbb{Z}$, let (n) be a nonzero proper ideal, and let $n = p_1^{e_1} \cdots p_t^{e_t}$ be the prime factorization. Then

$$(n) = (p_1^{e_1}) \cap \cdots \cap (p_t^{e_t})$$

is an irredundant primary decomposition.

(ii) Let $R = k[x, y]$, where k is a field. Define $Q_1 = (x)$ and $Q_2 = (x, y)^2$. Note that Q_1 is prime, and hence Q_1 is P_1-primary for $P_1 = Q_1$. Also, $P_2 = (x, y)$ is a maximal ideal, and so $Q_2 = P_2^2$ is P_2-primary, by Proposition 6.113. Define $I = Q_1 \cap Q_2$. This primary decompostion of I is irredundant. The associated primes of I are thus $\{P_1, P_2\}$. ◂

There is a second uniqueness result that describes a *normalized* primary decomposition, but we precede it by a lemma.

Lemma 6.118. *If P is a prime ideal and $Q_1, \ldots, Q_n$ are P-primary ideals, then $Q_1 \cap \cdots \cap Q_n$ is also a P-primary ideal.*

Proof. We verify that the three items in the hypothesis of Proposition 6.111 hold for $I = Q_1 \cap \cdots \cap Q_n$. Clearly, $I \subseteq P$. Second, if $b \in P$, then $b^{m_i} \in Q_i$ for all i, because Q_i is P-primary. Hence, $b^m \in I$, where $m = \max\{m_1, \ldots, m_n\}$. Finally, assume that $ab \in I$. If $a \notin I$, then $a \notin Q_i$ for some i. As Q_i is P-primary, $ab \in I \subseteq Q_i$ and $a \notin Q_i$ imply $b \in P$. Therefore, I is P-primary. •

Definition. A primary decomposition $I = Q_1 \cap \ldots \cap Q_r$ is ***normal*** if it is irredundant and if all the prime ideals $P_i = \sqrt{Q_i}$ are distinct.

Corollary 6.119. *If R is a noetherian ring, then every proper ideal in R has a normal primary decomposition.*

Proof. By Theorem 6.115, every proper ideal I has a primary decomposition, say,

$$I = Q_1 \cap \cdots \cap Q_r,$$

where Q_i is P_i-primary. If $P_r = P_i$ for some $i < r$, then Q_i and Q_r can be replaced by $Q' = Q_i \cap Q_r$, which is primary, by Lemma 6.118. Iterating, we eventually arrive at a primary decomposition with all prime ideals distinct. If this decomposition is not irredundant, remove primary ideals from it, one at a time, to obtain a normal primary decomposition. •

Definition. If $I = Q_1 \cap \cdots \cap Q_r$ is a normal primary decomposition, then the minimal prime ideals $P_i = \sqrt{Q_i}$ are called ***isolated*** prime ideals; the other prime ideals, if any, are called ***embedded***.

In Example 6.117(ii), we gave an irredundant primary decomposition of $I = (x) \cap (x, y)^2$ in $k[x, y]$, where k is a field. The associated primes are (x) and (x, y), so that (x) is an isolated prime and (x, y) is an embedded prime.

Definition. A prime ideal P is ***minimal*** over an ideal I if $I \subseteq P$ and there is no prime ideal P' with $I \subseteq P' \subsetneq P$.

Corollary 6.120. *Let I be an ideal in a noetherian ring R.*

(i) *Any two normal primary decompositions of I have the same set of isolated prime ideals, and so the isolated prime ideals are uniquely determined by I.*

(ii) *I has only finitely many minimal prime ideals.*

(iii) *A noetherian ring has only finitely many minimal prime ideals.*

Proof. (i) Let $I = Q_1 \cap \cdots \cap Q_n$ be a normal primary decomposition. If P is any prime ideal containing I, then

$$P \supseteq I = Q_1 \cap \cdots \cap Q_n \supseteq Q_1 \cdots Q_n.$$

Now $P \supseteq Q_i$ for some i, by Proposition 6.13, and so $P \supseteq \sqrt{Q_i} = P_i$. In other words, any prime ideal containing I must contain an isolated associated prime ideal. Hence, the isolated primes are the minimal elements in the set of associated primes of I; by Theorem 6.116, they are uniquely determined by I.

(ii) As in part (i), any prime ideal P containing I must contain an isolated prime of I. Hence, if P is minimal over I, then P must equal an isolated prime ideal of I. The result follows, for I has only finitely many isolated prime ideals.

(iii) This follows from part (ii) taking $I = \{0\}$. •

Here are some natural problems arising as these ideas are investigated further. First, what is the dimension of a variety? There are several candidates, and it turns out that prime ideals are the key. If V is a variety, then its dimension is the length of a longest chain of prime ideals in its coordinate ring $k[V]$ (which, by the correspondence theorem, is the length of a longest chain of prime ideals above $\mathrm{Id}(V)$ in $k[X]$).

It turns out to be more convenient to work in a larger ***projective space*** arising from k^n by adjoining a "hyperplane at infinity." For example, a projective plane arises from the usual plane by adjoining a line at infinity (it is the "horizon" where all parallel lines meet). To distinguish it from projective space, k^n is called ***affine space***, for it consists of the "finite points"—that is, not the points at infinity. If we study varieties in projective space, now defined as zeros of a set of *homogeneous* polynomials, then it is often the case that many separate affine cases become part of one simpler projective formula. For example, define the $\deg(C)$ to be the largest number of points in $C \cap \ell$, where ℓ is a line. If $C = \mathrm{Var}(f)$ is a curve arising from a polynomial of degree d, we want $\deg(C) = d$, but there are several problems here. First, we must demand that the coefficient field be algebraically closed, lest $\mathrm{Var}(f) = \varnothing$ cause a problem. Second, there may be multiple roots, and so some intersections may have to be counted with a certain *multiplicity*. *Bézout's theorem* states that if C and C' are two curves, then $|C \cap C'| = \deg(C)\deg(C')$. This formula holds in projective space, but it can be false in affine varieties. Defining multiplicities for intersections of higher-dimensional varieties is very subtle.

Finally, there is a deep analogy between differentiable manifolds and varieties. A *manifold* is a subspace of $\mathbb{R}^n$ that is a union of open replicas of euclidean space. For example, a torus T (i.e., a doughnut) is a subspace of $\mathbb{R}^3$, and each point of T has a neighborhood looking like an open disk (which is homeomorphic to the plane). We say that T is "locally euclidean"; it is obtained by gluing copies of $\mathbb{R}^2$ together in a coherent way. That a manifold is differentiable says there is a tangent space at each of its points. A variety V can be viewed as its coordinate ring $k[V]$, and neighborhoods of its points can be described "locally", using what is called a *sheaf* of local rings. If we "glue" sheaves together along open subsets having isomorphic sheaves of local rings, we obtain a *scheme*, and schemes seem to be the best way to study varieties. Two of the most prominent mathematicians involved in this circle of ideas are A. Grothendieck and J.-P. Serre.

EXERCISES

6.60 Prove that every algebraically closed field is infinite.

6.61 Prove that if an element a in a commutative ring R is nilpotent, then $1 + a$ is a unit.
Hint. The power series for $1/(1 + a)$ stops after a finite number of terms because a is nilpotent.

6.62 If I is an ideal in a commutative ring R, prove that its radical, $\sqrt{I}$, is an ideal.
Hint. If $f^r \in I$ and $g^s \in I$, prove that $(f + g)^{r+s} \in I$.

6.63 If R is a commutative ring, then its ***nilradical*** $\mathrm{nil}(R)$ is defined to be the intersection of all the prime ideals in R. Prove that $\mathrm{nil}(R)$ is the set of all the nilpotent elements in R:

$$\mathrm{nil}(R) = \{r \in R : r^m = 0 \text{ for some } m \geq 1\}.$$

Hint. If $r \in R$ is not nilpotent, use Exercise 6.9 on page 325 to show that there is some prime ideal not containing r.

6.64 (i) Show that $x^2 + y^2$ is irreducible in $\mathbb{R}[x, y]$, and conclude that $(x^2 + y^2)$ is a prime, hence radical, ideal in $\mathbb{R}[x, y]$.

(ii) Prove that $\mathrm{Var}(x^2 + y^2) = \{(0, 0)\}$.

(iii) Prove that $\mathrm{Id}(\mathrm{Var}(x^2 + y^2)) \supsetneq (x^2 + y^2)$, and conclude that the radical ideal $(x^2 + y^2)$ in $\mathbb{R}[x, y]$ is not of the form $\mathrm{Id}(V)$ for some variety V. Conclude that the Nullstellensatz may fail in $k[X]$ if k is not algebraically closed.

(iv) Prove that $(x^2 + y^2) = (x + iy) \cap (x - iy)$ in $\mathbb{C}[x, y]$.

(v) Prove that $\mathrm{Id}(\mathrm{Var}(x^2 + y^2)) = (x^2 + y^2)$ in $\mathbb{C}[x, y]$.

6.65 Prove that if k is an (uncountable) algebraically closed field and $f_1, \ldots, f_t \in k[X]$, then $\mathrm{Var}(f_1, \ldots, f_t) = \varnothing$ if and only if there are $h_1, \ldots, h_t \in k[X]$ such that

$$1 = \sum_{i=1}^{t} h_i(X) f_i(X).$$

6.66 Let k be an (uncountable) algebraically closed field, and let $I = (f_1, \ldots, f_t) \subseteq k[X]$. If $g(X) \in k[X]$, prove that $g \in \sqrt{I} \subseteq k[X]$ if and only if $(f_1, \ldots, f_t, 1 - yg)$ is not a proper ideal in $k[X, y]$.

Hint. Use the Rabinowitch trick.

6.67 Let R be a commutative ring, and let $\mathrm{Spec}(R)$ denote the set of all the prime ideals in R. If I is an ideal in R, define

$$\overline{I} = \{\text{all the prime ideals in } R \text{ containing } I\}.$$

Prove the following:

(i) $\overline{\{0\}} = \mathrm{Spec}(R)$.

(ii) $\overline{R} = \varnothing$.

(iii) $\overline{\sum_\ell I_\ell} = \bigcap_\ell \overline{I_\ell}$.

(iv) $\overline{I \cap J} = \overline{IJ} = \overline{I} \cup \overline{J}$.

Conclude that $\mathrm{Spec}(R)$ is a topological space whose closed subsets are the ***Zariski closed sets***: those sets of of the form $\overline{I}$, where I varies over the ideals in R.

6.68 Prove that an ideal P in $\mathrm{Spec}(R)$ is closed (that is, the one-point set $\{P\}$ is a Zariski closed set) if and only if P is a maximal ideal.

6.69 If X and Y are topological spaces, then a function $g \colon X \to Y$ is ***continuous*** if, for each closed subset Q of Y, the inverse image $g^{-1}(Q)$ is a closed subset of X.

Let $f \colon R \to A$ be a ring homomorphism, and define $f^* \colon \mathrm{Spec}(A) \to \mathrm{Spec}(R)$ by $f^*(Q) = f^{-1}(Q)$, where Q is any prime ideal in A. Prove that f^* is a continuous function. [Recall that $f^{-1}(Q)$ is a prime ideal, by Exercise 6.5 on page 325.]

6.70 Prove that the function $\varphi \colon k^n \to \mathrm{Spec}(k[x_1, \ldots, x_n])$ [where k is an (uncountable) algebraically closed field], defined by $\varphi \colon (a_1, \ldots, a_n) \mapsto (x_1 - a_1, \ldots, x_n - a_n)$, is a continous injection (where both k^n and $\mathrm{Spec}(k[x_1, \ldots, x_n])$ are equipped with the Zariski topology; the Zariski topology on k^n was defined just after Proposition 6.93).

6.71 Prove that any descending chain

$$F_1 \supseteq F_2 \supseteq \cdots \supseteq F_m \supseteq F_{m+1} \supseteq \cdots$$

of closed sets in k^n stops; there is some t with $F_t = F_{t+1} = \cdots$.

6.72 If R is a commutative noetherian ring, prove that an ideal I in R is a radical ideal if and only if $I = P_1 \cap \cdots \cap P_r$, where the P_i are prime ideals.

6.73 Prove that there is an ideal I in a commutative ring R with I not primary and with $\sqrt{I}$ prime. **Hint.** Take $R = k[x, y]$, where k is a field, and $I = (x^2, xy)$.

6.74 Let $R = k[x, y]$, where k is a field, and let $I = (x^2, y)$. For each $a \in k$, prove that $I = (x) \cap (y + ax, x^2)$ is an irredundant primary decomposition. Conclude that the primary ideals in an irredundant primary decomposition of an ideal need not be unique.

6.6 GRÖBNER BASES

There is a canard that classical Greek philosophers were reluctant to perform experiments, preferring pure reason. Rather than looking in one's mouth and counting, for example, they would speculate about how many teeth a person needs, deciding that every man should have, say, 28 teeth. Young mathematicians also prefer pure reasoning, but they, too, should count teeth. Computations and algorithms are useful, if for no other reason than to serve as data from which we might conjecture theorems. In this light, consider the problem of finding the irreducible components of a variety $\mathrm{Var}(I)$; algebraically, this problem asks for the associated primes of I. The primary decomposition theorem says that we should seek primary ideals Q_i containing I, and the desired components are $\mathrm{Var}(\sqrt{Q_i})$. In the proof of Theorem 6.116, however, we saw that if $I = Q_i \cap \cdots \cap Q_r$ is an irredundant primary decomposition, where Q_i is P_i-primary, then $P_i = \sqrt{(I : c_i)}$, where $c_i \in \bigcap_{j \neq i} Q_j$ with $c \notin Q_i$. Taking an honest look at the teeth involves the following question. Given a set of generators of I, can we find generators of P_i explicitly? The difficulty lies in finding the elements c_i, for we will show, in this section, how to find generators of $\sqrt{(I : c)}$. Having made this point, we must also say that algorithms can do more than provide data in particular cases. For example, the euclidean algorithm is used in an essential way in proving that if K/k is a field extension, and if $f(x), g(x) \in k[x]$, then their gcd in $K[x]$ is equal to their gcd in $k[x]$.

Given two polynomials $f(x), g(x) \in k[x]$ with $g(x) \neq 0$, where k is a field, when is $g(x)$ a divisor of $f(x)$? The division algorithm gives unique polynomials $q(x), r(x) \in k[x]$ with

$$f(x) = q(x)g(x) + r(x),$$

where $r = 0$ or $\deg(r) < \deg(g)$, and $g \mid f$ if and only if the remainder $r = 0$. Let us look at this formula from a different point of view. To say that $g \mid f$ is to say that $f \in (g)$, the principal ideal generated by $g(x)$. Thus, the remainder r is the obstruction to f lying in this ideal; that is, $f \in (g)$ if and only if $r = 0$.

Consider a more general problem. Given polynomials

$$f(x), g_1(x), \ldots, g_m(x) \in k[x],$$

where k is a field, when is $d(x) = \gcd\{g_1(x), \ldots, g_m(x)\}$ a divisor of f? The euclidean algorithm finds d, and the division algorithm determines whether $d \mid f$. From another viewpoint, the two classical algorithms combine to give an algorithm determining whether $f \in (g_1, \ldots, g_m) = (d)$.

We now ask whether there is an algorithm in $k[x_1, \ldots, x_n] = k[X]$ to determine, given $f(X), g_1(X), \ldots, g_m(X) \in k[X]$, whether $f \in (g_1, \ldots, g_m)$. A generalized division algorithm in $k[X]$ should be an algorithm yielding

$$r(X), a_1(X), \ldots, a_m(X) \in k[X],$$

with $r(X)$ unique, such that

$$f = a_1 g_1 + \cdots + a_m g_m + r$$

and $f \in (g_1, \ldots, g_m)$ if and only if $r = 0$. Since $(g_1, \ldots, g_m)$ consists of all the linear combinations of the g's, such an algorithm would say that the remainder r is the obstruction to f lying in $(g_1, \ldots, g_m)$.

We are going to show that both the division algorithm and the euclidean algorithm can be extended to polynomials in several variables. Even though these results are elementary, they were discovered only recently, in 1965, by B. Buchberger. Algebra has always dealt with algorithms, but the power and beauty of the axiomatic method has dominated the subject ever since Cayley and Dedekind in the second half of the nineteenth century. After the invention of the transistor in 1948, high-speed calculation became a reality, and old complicated algorithms, as well as new ones, could be implemented; a higher order of computing had entered algebra. Most likely, the development of computer science is a major reason why generalizations of the classical algorithms, from polynomials in one variable to polynomials in several variables, are only now being discovered. This is a dramatic illustration of the impact of external ideas on mathematics.

Generalized Division Algorithm

The most important feature of the division algorithm in $k[x]$ is that the remainder $r(x)$ has small degree. Without the inequality $\deg(r) < \deg(g)$, the result would be virtually useless; after all, given any $Q(x) \in k[x]$, there is an equation

$$f(x) = Q(x)g(x) + [f(x) - Q(x)g(x)].$$

Now polynomials in several variables are sums of monomials $cx_1^{\alpha_1} \cdots x_n^{\alpha_n}$, where $c \in k$ and $\alpha_i \geq 0$ for all i. Here are two degrees that we can assign to a monomial.

Definition. The ***multidegree*** of a monomial $cx_1^{\alpha_1} \cdots x_n^{\alpha_n} \in k[x_1, \ldots, x_n]$, where $c \in k$ is nonzero and $\alpha_i \geq 0$ for all i, is the n-tuple $\alpha = (\alpha_1, \ldots, \alpha_n)$; its ***weight*** is the sum $|\alpha| = \alpha_1 + \cdots + \alpha_n$.

When dividing $f(x)$ by $g(x)$ in $k[x]$, we usually arrange the monomials in $f(x)$ in descending order, according to degree:

$$f(x) = c_n x^n + c_{n-1} x^{n-1} + \cdots + c_2 x^2 + c_1 x + c_0.$$

A polynomial in several variables,

$$f(X) = f(x_1, \ldots, x_n) = \sum c_{(\alpha_1, \ldots, \alpha_n)} x_1^{\alpha_1} \cdots x_n^{\alpha_n},$$

can be written more compactly as

$$f(X) = \sum_{\alpha} c_\alpha X^\alpha$$

if we abbreviate $(\alpha_1, \ldots, \alpha_n)$ to α and $x_1^{\alpha_1} \cdots x_n^{\alpha_n}$ to X^α. We will arrange the monomials involved in $f(X)$ in a reasonable way by ordering their multidegrees.

In Example 5.69(ii), we saw that $\mathbb{N}^n$, the set of all n-tuples $\alpha = (\alpha_1, \ldots, \alpha_n)$ of natural numbers, is a monoid under addition:

$$\alpha + \beta = (\alpha_1, \ldots, \alpha_n) + (\beta_1, \ldots, \beta_n) = (\alpha_1 + \beta_1, \ldots, \alpha_n + \beta_n).$$

This monoid operation is related to the multiplication of monomials:

$$X^\alpha X^\beta = X^{\alpha+\beta}.$$

Recall that a *partially ordered set* is a set X equipped with a relation $\preceq$ that is reflexive, antisymmetric, and transitive. Of course, we may write $x \prec y$ if $x \preceq y$ and $x \neq y$, and we may write $y \succeq x$ (or $y \succ x$) instead of $x \preceq y$ (or $x \prec y$). A partially ordered set X is *well-ordered* if every nonempty subset $S \subseteq X$ contains a smallest element; that is, there exists $s_0 \in S$ with $s_0 \preceq s$ for all $s \in S$. For example, the least integer axiom says that the natural numbers $\mathbb{N}$ with the usual inequality $\leq$ is well-ordered.

Proposition A.3 in the Appendix proves that every strictly decreasing sequence in a well-ordered set must be finite. This property of well-ordered sets can be used to show that an algorithm eventually stops. For example, in the proof of the division algorithm for polynomials in one variable, we associated a natural number to each step: the degree of a remainder. Moreover, if the algorithm does not stop at a given step, then the natural number associated to the next step—the degree of its remainder—is strictly smaller. Since the natural numbers are well-ordered by the usual inequality $\leq$, this strictly decreasing sequence of natural numbers must be finite; that is, the algorithm must stop after a finite number of steps.

We are interested in orderings of multidegrees that are compatible with multiplication of monomials—that is, with addition in the monoid $\mathbb{N}^n$.

Definition. A ***monomial order*** is a well-ordering of $\mathbb{N}^n$ such that

$$\alpha \preceq \beta \quad \text{implies} \quad \alpha + \gamma \preceq \beta + \gamma$$

for all $\alpha, \beta, \gamma \in \mathbb{N}^n$.

A monomial order will be used as follows. If $X = (x_1, \ldots, x_n)$, then we define $X^\alpha \preceq X^\beta$ in case $\alpha \preceq \beta$; that is, monomials are ordered according to their multidegrees.

Definition. If $\mathbb{N}^n$ is equipped with a monomial order, then every $f(X) \in k[X] = k[x_1, \ldots, x_n]$ can be written with its largest term first, followed by its other, smaller, terms in descending order:

$$f(X) = c_\alpha X^\alpha + \text{ lower terms.}$$

Define its ***leading term*** to be $\mathrm{LT}(f) = c_\alpha X^\alpha$ and its ***Degree*** to be $\mathrm{Deg}(f) = \alpha$. Call $f(X)$ ***monic*** if $\mathrm{LT}(f) = X^\alpha$; that is, if $c_\alpha = 1$.

Note that $\mathrm{Deg}(f)$ and $\mathrm{LT}(f)$ depend on the monomial order.

There are many examples of monomial orders, but we shall give only the two most popular ones.

Definition. The ***lexicographic order*** on $\mathbb{N}^n$ is defined by $\alpha \preceq_{\text{lex}} \beta$ if either $\alpha = \beta$ or the first nonzero coordinate in $\beta - \alpha$ is positive.[17]

The term *lexicographic* refers to the standard ordering of words in a dictionary. For example, the following German words are increasing in lexicographic order (the letters are ordered $a < b < c < \cdots < z$):

ausgehen
ausladen
auslagen
auslegen
bedeuten

If $\alpha \prec_{\text{lex}} \beta$, then they agree for the first $i - 1$ coordinates (for some $i \geq 1$), that is, $\alpha_1 = \beta_1, \ldots, \alpha_{i-1} = \beta_{i-1}$, and there is strict inequality: $\alpha_i < \beta_i$.

Proposition 6.121. *The lexicographic order $\preceq_{\text{lex}}$ is a monomial order on $\mathbb{N}^n$.*

Proof. First, we show that the lexicographic order is a partial order. The relation $\preceq_{\text{lex}}$ is reflexive, for its definition shows that $\alpha \preceq_{\text{lex}} \alpha$. To prove antisymmetry, assume that $\alpha \preceq_{\text{lex}} \beta$ and $\beta \preceq_{\text{lex}} \alpha$. If $\alpha \neq \beta$, there is a first coordinate, say the ith, where they disagree. For notation, we may assume that $\alpha_i < \beta_i$. But this contradicts $\beta \preceq_{\text{lex}} \alpha$.

[17] The difference $\beta - \alpha$ may not lie in $\mathbb{N}^n$, but it does lie in $\mathbb{Z}^n$.

To prove transitivity, suppose that $\alpha \prec_{\text{lex}} \beta$ and $\beta \prec_{\text{lex}} \gamma$ (it suffices to consider strict inequality). Now $\alpha_1 = \beta_1, \ldots, \alpha_{i-1} = \beta_{i-1}$ and $\alpha_i < \beta_i$. Let γ_p be the first coordinate with $\beta_p < \gamma_p$. If $p < i$, then

$$\gamma_1 = \beta_1 = \alpha_1, \ldots, \gamma_{p-1} = \beta_{p-1} = \alpha_{p-1},\ \alpha_p = \beta_p < \gamma_p;$$

if $p \geq i$, then

$$\gamma_1 = \beta_1 = \alpha_1, \ldots, \gamma_{i-1} = \beta_{i-1} = \alpha_{i-1},\ \alpha_i < \beta_i = \gamma_i.$$

In either case, the first nonzero coordinate of $\gamma - \alpha$ is positive; that is, $\alpha \prec_{\text{lex}} \gamma$.

Next, we show that the lexicographic order is a well-order. If S is a nonempty subset of $\mathbb{N}^n$, define

$$C_1 = \{\text{all first coordinates of } n\text{-tuples in } S\},$$

and define δ_1 to be the smallest number in C_1 (note that C_1 is a nonempty subset of the well-ordered set $\mathbb{N}$). Define

$$C_2 = \{\text{all second coordinates of } n\text{-tuples } (\delta_1, \alpha_2, \ldots, \alpha_n) \in S\}.$$

Since $C_2 \neq \varnothing$, it contains a smallest number, δ_2. Similarly, for all $i < n$, define C_{i+1} as all the $(i+1)$th coordinates of those n-tuples in S whose first i coordinates are $(\delta_1, \delta_2, \ldots, \delta_i)$, and define δ_{i+1} to be the smallest number in C_{i+1}. By construction, the n-tuple $\delta = (\delta_1, \delta_2, \ldots, \delta_n)$ lies in S; moreover, if $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_n) \in S$, then

$$\alpha - \delta = (\alpha_1 - \delta_1, \alpha_2 - \delta_2, \ldots, \alpha_n - \delta_n)$$

has its first nonzero coordinate, if any, positive, and so $\delta \prec_{\text{lex}} \alpha$. Therefore, the lexicographic order is a well-order.

Assume that $\alpha \preceq_{\text{lex}} \beta$; we claim that

$$\alpha + \gamma \preceq_{\text{lex}} \beta + \gamma$$

for all $\gamma \in \mathbb{N}$. If $\alpha = \beta$, then $\alpha + \gamma = \beta + \gamma$. If $\alpha \prec_{\text{lex}} \beta$, then the first nonzero coordinate of $\beta - \alpha$ is positive. But

$$(\beta + \gamma) - (\alpha + \gamma) = \beta - \alpha,$$

and so $\alpha + \gamma \prec_{\text{lex}} \beta + \gamma$. Therefore, $\preceq_{\text{lex}}$ is a monomial order. •

In the lexicographic order, $x_1 \succ x_2 \succ x_3 \succ \cdots$, for

$$(1, 0, \ldots, 0) \succ (0, 1, 0, \ldots, 0) \succ \cdots \succ (0, 0, \ldots, 1).$$

Any permutation of the variables $x_{\sigma(1)}, \ldots, x_{\sigma(n)}$ yields a different lexicographic order on $\mathbb{N}^n$.

Remark. If X is any well-ordered set with order $\preceq$, then the lexicographic order on X^n can be defined by $a = (a_1, \ldots, a_n) \preceq_{\text{lex}} b = (b_1, \ldots, b_n)$ in case $a = b$ or if they first disagree in the ith coordinate and $a_i \prec b_i$. It is a simple matter to generalize Proposition 6.121 by replacing $\mathbb{N}$ with X. ◂

In Lemma 5.70 we constructed, for any set X, a monoid $\mathcal{W}(X)$: its elements are the empty word together with all the words $x_1^{e_1} \cdots x_p^{e_p}$ on a set X, where $p \geq 1$ and $e_i = \pm 1$ for all i; its operation is juxtaposition. In contrast to $\mathbb{N}^n$, in which all words have length n, the monoid $\mathcal{W}(X)$ has words of different lengths. Of more interest here is the submonoid $\mathcal{W}^+(X)$ of $\mathcal{W}(X)$ consisting of all the "positive" words on X:

$$\mathcal{W}^+(X) = \{x_1 \cdots x_p \in \mathcal{W}(X) : x_i \in X \text{ and } p \geq 0\}.$$

Corollary 6.122. *If X is a well-ordered set, then $\mathcal{W}^+(X)$ is well-ordered in the lexicographic order* (*which we also denote by $\preceq_{\text{lex}}$*).

Proof. We will only give a careful definition of the lexicographic order here; the proof that it is a well-order is left to the reader. First, define $1 \preceq_{\text{lex}} w$ for all $w \in \mathcal{W}^+(X)$. Next, given words $u = x_1 \cdots x_p$ and $v = y_1 \cdots y_q$ in $\mathcal{W}^+(X)$, make them the same length by adjoining 1's at the end of the shorter word, and rename them u' and v' in $\mathcal{W}^+(X)$. If $m \geq \max\{p, q\}$, we may regard $u', v', \in X^m$, and we define $u \preceq_{\text{lex}} v$ if $u' \preceq_{\text{lex}} v'$ in X^m. (This is the word order commonly used in dictionaries, where a blank precedes any letter: for example, *muse* precedes *museum*.) •

Lemma 6.123. *Given a monomial order on $\mathbb{N}^n$, any sequence of steps of the form $f(X) \to f(X) - c_\beta X^\beta + g(X)$, where $c_\beta X^\beta$ is a nonzero term of $f(X)$ and* $\text{Deg}(g) \prec \beta$, *must be finite.*

Proof. Each polynomial

$$f(X) = \sum_\alpha c_\alpha X^\alpha \in k[X] = k[x_1, \ldots, x_n]$$

can be written with the multidegrees of its terms in descending order: $\alpha_1 \succ \alpha_2 \succ \cdots \succ \alpha_p$. Define

$$\text{multiword}(f) = \alpha_1 \cdots \alpha_p \in \mathcal{W}^+(\mathbb{N}^n).$$

Let $c_\beta X^\beta$ be a nonzero term in $f(X)$, let $g(X) \in k[X]$ have $\text{Deg}(g) \prec \beta$, and write

$$f(X) = h(X) + c_\beta X^\beta + \ell(X),$$

where $h(X)$ is the sum of all terms in $f(X)$ of multidegree $\succ \beta$ and $\ell(X)$ is the sum of all terms in $f(X)$ of multidegree $\prec \beta$. We claim that

$$\begin{aligned} \text{multiword}(f(X) - c_\beta X^\beta + g(X)) &\preceq_{\text{lex}} \text{multiword}(h + \ell + g) \\ &\prec_{\text{lex}} \text{multiword}(f) \text{ in } \mathcal{W}^+(X). \end{aligned}$$

The sum of the terms in $f(X) - c_\beta X^\beta + g(X)$ with multidegree $\succ \beta$ is $h(X)$, while the sum of the lower terms is $\ell(X) + g(X)$. But $\mathrm{Deg}(\ell + g) \prec \beta$, by Exercise 6.79 on page 410. Therefore, the initial terms of $f(X)$ and $f(X) - c_\beta X^\beta + g(X)$ agree, while the next term of $f(X) - c_\beta X^\beta + g(X)$ has multidegree $\prec \beta$, and this proves the claim. Since $\mathcal{W}^+(\mathbb{N}^n)$ is well-ordered, it follows that any sequence of steps of the form $f(X) \to f(X) - c_\beta X^\beta + g(X)$ must be finite. •

Here is the second popular monomial order. Recall that if $\alpha = (\alpha_1, \ldots, \alpha_n) \in \mathbb{N}^n$, then $|\alpha| = \alpha_1 + \cdots + \alpha_n$ denotes its weight.

Definition. The ***degree-lexicographic order*** on $\mathbb{N}^n$ is defined by $\alpha \preceq_{\mathrm{dlex}} \beta$ if either $\alpha = \beta$ or

$$|\alpha| = \sum_{i=1}^{n} \alpha_i < \sum_{i=1}^{n} \beta_i = |\beta|,$$

or, if $|\alpha| = |\beta|$, then the first nonzero coordinate in $\beta - \alpha$ is positive.

In other words, given $\alpha = (\alpha_1, \ldots, \alpha_n)$ and $\beta = (\beta_1, \ldots, \beta_n)$, first check weights: if $|\alpha| < |\beta|$, then $\alpha \preceq_{\mathrm{dlex}} \beta$; if there is a tie, that is, if α and β have the same weight, then order them lexicographically. For example, $(1, 2, 3, 0) \prec_{\mathrm{dlex}} (0, 2, 5, 0)$ and $(1, 2, 3, 4) \prec_{\mathrm{dlex}} (1, 2, 5, 2)$.

Proposition 6.124. *The degree-lexicographic order $\preceq_{\mathrm{dlex}}$ is a monomial order on $\mathbb{N}^n$.*

Proof. It is routine to show that $\preceq_{\mathrm{dlex}}$ is a partial order on $\mathbb{N}^n$. To see that it is a well-order, let S be a nonempty subset of $\mathbb{N}^n$. The weights of elements in S form a nonempty subset of $\mathbb{N}$, and so there is a smallest such, say, t. The nonempty subset of all $\alpha \in S$ having weight t has a smallest element, because the degree-lexicographic order $\preceq_{\mathrm{dlex}}$ coincides with the lexicographic order $\preceq_{\mathrm{lex}}$ on this subset. Therefore, there is a smallest element in S in the degree-lexicographic order.

Assume that $\alpha \preceq_{\mathrm{dlex}} \beta$ and $\gamma \in \mathbb{N}^n$. Now $|\alpha + \gamma| = |\alpha| + |\gamma|$, so that $|\alpha| = |\beta|$ implies $|\alpha + \gamma| = |\beta + \gamma|$ and $|\alpha| < |\beta|$ implies $|\alpha + \gamma| < |\beta + \gamma|$; in the latter case, Proposition 6.121 shows that $\alpha + \gamma \preceq_{\mathrm{dlex}} \beta + \gamma$. •

The next proposition shows, with respect to a monomial order, that polynomials in several variables behave like polynomials in a single variable.

Proposition 6.125. *Let $\preceq$ be a monomial order on $\mathbb{N}^n$, and let $f(X), g(X), h(X) \in k[X] = k[x_1, \ldots, x_n]$, where k is a field.*

(i) *If* $\mathrm{Deg}(f) = \mathrm{Deg}(g)$, *then* $\mathrm{LT}(g) \mid \mathrm{LT}(f)$.

(ii) $\mathrm{LT}(hg) = \mathrm{LT}(h)\mathrm{LT}(g)$.

(iii) *If* $\mathrm{Deg}(f) = \mathrm{Deg}(hg)$, *then* $\mathrm{LT}(g) \mid \mathrm{LT}(f)$.

Proof. (i) If $\mathrm{Deg}(f) = \alpha = \mathrm{Deg}(g)$, then $\mathrm{LT}(f) = cX^\alpha$ and $\mathrm{LT}(g) = dX^\alpha$. Hence, $\mathrm{LT}(g) \mid \mathrm{LT}(f)$, because $c \neq 0$ and so c is a unit in k [note that $\mathrm{LT}(f) \mid \mathrm{LT}(g)$ as well].

(ii) Let $h(X) = cX^\gamma + \text{lower terms}$ and let $g(X) = bX^\beta + \text{lower terms}$, so that $\mathrm{LT}(h) = cX^\gamma$ and $\mathrm{LT}(g) = bX^\beta$. Clearly, $cbX^{\gamma+\beta}$ is a nonzero term of $h(X)g(X)$. To see that it is the leading term, let $c_\mu X^\mu$ be a term of $h(X)$ with $\mu \preceq \gamma$, and let $b_\nu X^\nu$ be a term of $g(X)$ with $\nu \preceq \beta$ (with at least one strict inquality). Now $\mathrm{Deg}(c_\mu X^\mu b_\nu X^\nu) = \mu + \nu$; since $\preceq$ is a monomial order, we have $\mu + \nu \prec \gamma + \nu \prec \gamma + \beta$. Thus, $cbX^{\gamma+\beta}$ is the term in $h(X)g(X)$ with largest multidegree.

(iii) Since $\mathrm{Deg}(f) = \mathrm{Deg}(hg)$, part (i) gives $\mathrm{LT}(hg) \mid \mathrm{LT}(f)$, and $\mathrm{LT}(h)\mathrm{LT}(g) = \mathrm{LT}(hg)$, by part (ii); hence, $\mathrm{LT}(g) \mid \mathrm{LT}(f)$. •

Definition. Let $\preceq$ be a monomial order on $\mathbb{N}^n$ and let $f(X), g(X) \in k[X]$, where $k[X] = k[x_1, \ldots, x_n]$. If there is a nonzero term $c_\beta X^\beta$ in $f(X)$ with $\mathrm{LT}(g) \mid c_\beta X^\beta$ and

$$h(X) = f(X) - \frac{c_\beta X^\beta}{\mathrm{LT}(g)} g(X),$$

then the ***reduction*** $f \xrightarrow{g} h$ is the replacement of f by h.

Reduction is precisely the usual step involved in long division of polynomials in one variable. Of course, a special case of reduction is when $c_\beta X^\beta = \mathrm{LT}(f)$.

Proposition 6.126. *Let $\preceq$ be a monomial order on $\mathbb{N}^n$, let $f(X), g(X) \in k[X] = k[x_1, \ldots, x_n]$, and assume that $f \xrightarrow{g} h$; that is, there is a nonzero term $c_\beta X^\beta$ in $f(X)$ with* $\mathrm{LT}(g) \mid c_\beta X^\beta$ *and* $h(X) = f(X) - \frac{c_\beta X^\beta}{\mathrm{LT}(g)} g(X)$. *Then*

$$\mathrm{Deg}\left(\frac{c_\beta X^\beta}{\mathrm{LT}(g)} g(X)\right) \preceq \mathrm{Deg}(f).$$

Moreover, if $\beta = \mathrm{Deg}(f)$ [*i.e., if* $c_\beta X^\beta = \mathrm{LT}(f)$], *then either*

$$h(X) = 0 \quad \text{or} \quad \mathrm{Deg}(h) \prec \mathrm{Deg}(f),$$

and if $\beta \prec \mathrm{Deg}(f)$, *then* $\mathrm{Deg}(h) = \mathrm{Deg}(f)$.

Proof. Let us write

$$f(X) = \mathrm{LT}(f) + c_\kappa X^\kappa + \text{lower terms},$$

where $c_\kappa X^\kappa = \mathrm{LT}(f - \mathrm{LT}(f))$; since $c_\beta X^\beta$ is a term of $f(X)$, we have $\beta \preceq \mathrm{Deg}(f)$. Similarly, if $\mathrm{LT}(g) = a_\gamma X^\gamma$, so that $\mathrm{Deg}(g) = \gamma$, let us write

$$g(X) = a_\gamma X^\gamma + a_\lambda X^\lambda + \text{lower terms},$$

where $a_\lambda X^\lambda = \mathrm{LT}(g - \mathrm{LT}(g))$. Hence,

$$\begin{aligned} h(X) &= f(X) - \frac{c_\beta X^\beta}{\mathrm{LT}(g)} g(X) \\ &= f(X) - \frac{c_\beta X^\beta}{\mathrm{LT}(g)} \Big[\mathrm{LT}(g) + a_\lambda X^\lambda + \cdots \Big] \\ &= \Big[f(X) - c_\beta X^\beta \Big] - \frac{c_\beta X^\beta}{\mathrm{LT}(g)} \Big[a_\lambda X^\lambda + \cdots \Big]. \end{aligned}$$

Now $\mathrm{LT}(g) \mid c_\beta X^\beta$ says that $\beta - \gamma \in \mathbb{N}^n$. We claim that

$$\mathrm{Deg}\left(- \frac{c_\beta X^\beta}{\mathrm{LT}(g)} \Big[a_\lambda X^\lambda + \cdots \Big] \right) = \lambda + \beta - \gamma;$$

that is, $\lambda + \beta - \gamma = \mathrm{Deg}\left(- \frac{c_\beta X^\beta}{\mathrm{LT}(g)} a_\lambda X^\lambda \right)$ is the largest multidegree occurring. Suppose that $a_\eta X^\eta$ is a lower term in $g(X)$ (i.e., $\eta \prec \lambda$); since $\preceq$ is a monomial order,

$$\eta + (\beta - \gamma) \prec \gamma + (\lambda - \gamma) = \lambda.$$

Now $\lambda \prec \gamma$ implies $\lambda + (\beta - \gamma) \prec \gamma + (\beta - \gamma) = \beta$, and so

$$\mathrm{Deg}\left(- \Big[\frac{c_\beta X^\beta}{\mathrm{LT}(g)} \Big] g(X) \right) \prec \beta \preceq \mathrm{Deg}(f). \tag{6}$$

Therefore, if $h(X) \neq 0$, then Exercise 6.79 on page 410 gives

$$\mathrm{Deg}(h) \preceq \max \left\{ \mathrm{Deg}\left(f(X) - c_\beta X^\beta \right), \mathrm{Deg}\left(- \left[\frac{c_\beta X^\beta}{\mathrm{LT}(g)} \right] g(X) \right) \right\}.$$

Now if $\beta = \mathrm{Deg}(f)$, then $c_\beta X^\beta = \mathrm{LT}(f)$,

$$f(X) - c_\beta X^\beta = f(X) - \mathrm{LT}(f) = c_\kappa X^\kappa + \text{lower terms},$$

and, hence, $\mathrm{Deg}(f(X) - c_\beta X^\beta) = \kappa \prec \mathrm{Deg}(f)$ in this case. If $\beta \prec \mathrm{Deg}(f)$, then $\mathrm{Deg}(f(X) - c_\beta X^\beta) = \mathrm{Deg}(f)$, while $\mathrm{Deg}\left(- \Big[\frac{c_\beta X^\beta}{\mathrm{LT}(g)} \Big] g(X) \right) \prec \mathrm{Deg}(f)$, by Eq. (6), and so $\mathrm{Deg}(h) = \mathrm{Deg}(f)$ in this case.

The last inequality is clear, for

$$\frac{c_\beta X^\beta}{\mathrm{LT}(g)} g(X) = c_\beta X^\beta + \frac{c_\beta X^\beta}{\mathrm{LT}(g)} \Big[a_\lambda X^\lambda + \cdots \Big].$$

Since the latter part of the polynomial has Degree $\lambda + \beta - \gamma \prec \beta$, we see that

$$\mathrm{Deg}\left(\frac{c_\beta X^\beta}{\mathrm{LT}(g)} g(X) \right) = \beta \preceq \mathrm{Deg}(f). \quad \bullet$$

Definition. Let $\{g_1(X), \ldots, g_m(X)\}$ be a set of polynomials in $k[X]$. A polynomial $r(X)$ is ***reduced mod*** $\{g_1, \ldots, g_m\}$ if either $r(X) = 0$ or no $\mathrm{LT}(g_i)$ divides any nonzero term of $r(X)$.

Here is the division algorithm for polynomials in several variables. Because the algorithm requires the "divisor polynomials" $\{g_1, \ldots, g_m\}$ to be used in a specific order (after all, an algorithm must give explicit directions), we will be using an m-tuple of polynomials instead of a subset of polynomials. We denote the m-tuple whose ith entry is g_i by $[g_1, \ldots, g_m]$, because the usual notation $(g_1, \ldots, g_m)$ would be confused with the ideal $(g_1, \ldots, g_m)$ generated by the g_i.

Theorem 6.127 (Division Algorithm in $k[X]$). *Let $\preceq$ be a monomial order on $\mathbb{N}^n$, and let $k[X] = k[x_1, \ldots, x_n]$. If $f(X) \in k[X]$ and $G = [g_1(X), \ldots, g_m(X)]$ is an m-tuple of polynomials in $k[X]$, then there is an algorithm giving polynomials $r(X), a_1(X), \ldots, a_m(X) \in k[X]$ with*

$$f = a_1 g_1 + \cdots + a_m g_m + r,$$

where r is reduced mod $\{g_1, \ldots, g_m\}$*, and*

$$\mathrm{Deg}(a_i g_i) \preceq \mathrm{Deg}(f) \quad \textit{for all } i.$$

Proof. Once a monomial order is chosen, so that leading terms are defined, the algorithm is a straightforward generalization of the division algorithm in one variable. First, reduce mod g_1 as many times as possible, then reduce mod g_2 as many times as possible, and then reduce again mod g_1; more generally, once a polynomial is reduced mod $[g_1, \ldots, g_i]$ for any i, then reduce mod $[g_1 \ldots, g_i, g_{i+1}]$. Here is a pseudocode describing the algorithm more precisely.

Input: $f(X) = \sum_\beta c_\beta X^\beta, \quad [g_1, \ldots, g_m]$
Output: $r, a_1, \ldots, a_m$
$r := f; \quad a_i := 0$
WHILE f is not reduced mod $\{g_1, \ldots, g_m\}$ DO
  select smallest i with $\mathrm{LT}(g_i) \mid c_\beta X^\beta$ for some β
  $f - [c_\beta X^\beta / \mathrm{LT}(g_i)] g_i := f$
  $a_i + [c_\beta X^\beta / \mathrm{LT}(g_i)] := a_i$
END WHILE

At each step $h_j \xrightarrow{g_i} h_{j+1}$ of the algorithm, we have

$$\mathrm{multiword}(h_j) \succ_{\mathrm{lex}} \mathrm{multiword}(h_{j+1})$$

in $\mathcal{W}^+(\mathbb{N}^n)$, by Lemma 6.123, and so the algorithm does stop, because $\prec_{\mathrm{lex}}$ is a well-order on $\mathcal{W}^+(\mathbb{N}^n)$. Obviously, the output $r(X)$ is reduced mod $\{g_1, \ldots, g_m\}$, for if it has a term divisible by some $\mathrm{LT}(g_i)$, then one further reduction is possible.

Finally, each term of $a_i(X)$ has the form $c_\beta X^\beta / \mathrm{LT}(g_i)$ for some intermediate output $h(X)$ (as we see in the pseudocode). It now follows from Proposition 6.126 that $\mathrm{Deg}(a_i g_i) \preceq \mathrm{Deg}(f)$. •

Definition. Given a monomial order on $\mathbb{N}^n$, a polynomial $f(X) \in k[X]$, and an m-tuple $G = [g_1, \ldots, g_m]$, we call the output $r(X)$ of the division algorithm the ***remainder of*** $f(X)$ ***mod*** G.

Note that the remainder r of f mod G is reduced mod $\{g_1, \ldots, g_m\}$ and $f - r \in I = (g_1, \ldots, g_m)$. The algorithm requires that G be an m-tuple, because of the command

$$\text{select smallest } i \text{ with } \mathrm{LT}(g_i) \mid c_\beta X^\beta \text{ for some } \beta$$

specifying the order of reductions.

The next example shows that the remainder may depend not only on the set of polynomials $\{g_1, \ldots, g_m\}$ but also on the ordering of the coordinates in the m-tuple $G = [g_1, \ldots, g_m]$. That is, if $\sigma \in S_m$ is a permutation and $G_\sigma = [g_{\sigma(1)}, \ldots, g_{\sigma(m)}]$, then the remainder r_σ of f mod G_σ may not be the same as the remainder r of f mod G. Even worse, it is possible that $r \neq 0$ and $r_\sigma = 0$, so that the remainder mod G is not the obstruction to f being in the ideal $(g_1, \ldots, g_m)$. We illustrate this phenomenon in the next example, and we will deal with it in the next subsection.

Example 6.128.
Let $f(x, y, z) = x^2y^2 + xy$, and let $G = [g_1, g_2, g_3]$, where

$$\begin{aligned} g_1 &= y^2 + z^2 \\ g_2 &= x^2y + yz \\ g_3 &= z^3 + xy. \end{aligned}$$

We use the degree-lexicographic order on $\mathbb{N}^3$. Now $y^2 = \mathrm{LT}(g_1) \mid \mathrm{LT}(f) = x^2y^2$, and so $f \xrightarrow{g_1} h$, where

$$h = f - \frac{x^2y^2}{y^2}(y^2 + z^2) = -x^2z^2 + xy.$$

The polynomial $-x^2z^2 + xy$ is reduced mod G, because neither $-x^2z^2$ nor xy is divisible by any of the leading terms $\mathrm{LT}(g_1) = y^2$, $\mathrm{LT}(g_2) = x^2y$, or $\mathrm{LT}(g_3) = z^3$.

Let us now apply the division algorithm using the 3-tuple $G' = [g_2, g_1, g_3]$. The first reduction gives $f \xrightarrow{g_2} h'$, where

$$h' = f - \frac{x^2y^2}{x^2y}(x^2y + yz) = -y^2z + xy.$$

Now h' is not reduced, and reducing mod g_1 gives

$$h' - \frac{-y^2z}{y^2}(y^2 + z^2) = z^3 + xy.$$

But $z^3 + xy = g_3$, and so $z^3 + xy \xrightarrow{g_3} 0$. Thus, the remainder depends on the ordering of the divisor polynomials g_i in the m-tuple.

For a simpler example of different remainders (but with neither remainder being 0), see Exercise 6.78. ◄

EXERCISES

6.75 (i) Let $(X, \preceq)$ and $(Y, \preceq')$ be well-ordered sets, where X and Y are disjoint. Define a binary relation $\leq$ on $X \cup Y$ by

$$\begin{aligned} x_1 \leq x_2 &\quad \text{if } x_1, x_2 \in X \text{ and } x_1 \preceq x_2,\\ y_1 \leq y_2 &\quad \text{if } y_1, y_2 \in Y \text{ and } y_1 \preceq' y_2,\\ x \leq y &\quad \text{if } x \in X \text{ and } y \in Y. \end{aligned}$$

Prove that $(X \cup Y, \leq)$ is a well-ordered set.

(ii) If $r \leq n$, we may regard $\mathbb{N}^r$ as the subset of $\mathbb{N}^n$ consisting of all n-tuples of the form $(n_1, \ldots, n_r, 0, \ldots, 0)$, where $n_i \in \mathbb{N}$ for all $i \leq r$. Prove that there exists a monomial order on $\mathbb{N}^n$ in which $a \prec b$ whenever $\alpha \in \mathbb{N}^r$ and $\beta \in \mathbb{N}^n - \mathbb{N}^r$.

Hint. Consider the lex order on $k[x_1, \ldots, x_n]$ in which $x_1 \prec x_2 \prec \cdots \prec x_n$.

6.76 (i) Write the first 10 monic monomials in $k[x, y]$ in lexicographic order and in degree-lexicographic order.

(ii) Write all the monic monomials in $k[x, y, z]$ of weight at most 2 in lexicographic order and in degree-lexicographic order.

6.77 Give an example of a well-ordered set X containing an element u having infinitely many predecessors; that is, $\{x \in X : x \preceq u\}$ is infinite.

6.78 Let $G = [x - y, x - z]$ and $G' = [x - z, x - y]$. Show that the remainder of $x \bmod G$ (in degree-lexicographic order) is distinct from the remainder of $x \bmod G'$.

6.79 Let $\preceq$ be a monomial order on $\mathbb{N}^n$, and let $f(X), g(X) \in k[X] = k[x_1, \ldots, x_n]$ be nonzero.

(i) Prove that if $f + g \neq 0$, then $\mathrm{Deg}(f + g) \preceq \max\{\mathrm{Deg}(f), \mathrm{Deg}(g)\}$, and that strict inequality can occur only if $\mathrm{Deg}(f) = \mathrm{Deg}(g)$.

(ii) Prove that $\mathrm{Deg}(fg) = \mathrm{Deg}(f) + \mathrm{Deg}(g)$, and $\mathrm{Deg}(f^m) = m\,\mathrm{Deg}(f)$ for all $m \geq 1$.

6.80 Use the degree-lexicographic order in this exercise.

(i) Find the remainder of $x^7y^2 + x^3y^2 - y + 1 \bmod [xy^2 - x, x - y^3]$.

(ii) Find the remainder of $x^7y^2 + x^3y^2 - y + 1 \bmod [x - y^3, xy^2 - x]$.

6.81 Use the degree-lexicographic order in this exercise.

(i) Find the remainder of $x^2y + xy^2 + y^2 \bmod [y^2 - 1, xy - 1]$.

(ii) Find the remainder of $x^2y + xy^2 + y^2 \bmod [xy - 1, y^2 - 1]$.

6.82 Let $c_\alpha X^\alpha$ be a nonzero monomial, and let $f(X), g(X) \in k[X]$ be polynomials none of whose terms is divisible by $c_\alpha X^\alpha$. Prove that none of the terms of $f(X) - g(X)$ is divisible by $c_\alpha X^\alpha$.

6.83 An ideal I in $k[X]$ that is generated by monomials, say, $I = (X^{\alpha(1)}, \ldots, X^{\alpha(q)})$, is called a ***monomial ideal***.

(i) Prove that $f(X) \in I$ if and only if each term of $f(X)$ is divisible by some $X^{\alpha(i)}$.

(ii) Prove that if $G = [g_1, \ldots, g_m]$ and r is reduced mod G, then r does not lie in the monomial ideal $(\mathrm{LT}(g_1), \ldots, \mathrm{LT}(g_m))$.

6.84 Let $f(X) = \sum_\alpha c_\alpha X^\alpha \in k[X]$ be symmetric, where k is a field and $X = (x_1, \ldots, x_n)$. Assume that $\mathbb{N}^n$ is equipped with the degree-lexicographic order and that $\mathrm{Deg}(f) = \beta = (\beta_1, \ldots, \beta_n)$.

(i) Prove that if $c_\alpha x_1^{\alpha_1} \cdots x_n^{\alpha_n}$ occurs with nonzero coefficient c_α, then every monomial $x_{\sigma 1}^{\alpha_1} \cdots x_{\sigma n}^{\alpha_n}$ also occurs in $f(X)$ with nonzero coefficient, where $\sigma \in S_n$.

(ii) Prove that $\beta_1 \geq \beta_2 \geq \cdots \geq \beta_n$.

(iii) If $e_1, \ldots, e_n$ are the elementary symmetric polynomials, prove that

$$\mathrm{Deg}(e_i) = (1, \ldots, 1, 0, \ldots, 0),$$

where there are i 1's.

(iv) Let $(\gamma_1, \ldots, \gamma_n) = (\beta_1 - \beta_2, \beta_2 - \beta_3, \ldots, \beta_{n-1} - \beta_n, \beta_n)$. Prove that if $g(x_1, \ldots, x_n) = x_1^{\gamma_1} \cdots x_n^{\gamma_n}$, then $g(e_1, \ldots, e_n)$ is symmetric and $\mathrm{Deg}(g) = \beta$.

(v) ***Fundamental Theorem of Symmetric Polynomials.*** Prove that if k is a field, then every symmetric polynomial $f(X) \in k[X]$ is a *polynomial* in the elementary symmetric functions $e_1, \ldots, e_n$. (Compare with Theorem 4.37.)

Hint. Prove that $h(X) = f(X) - c_\beta g(e_1, \ldots, e_n)$ is symmetric, and that $\mathrm{Deg}(h) < \beta$.

Buchberger's Algorithm

For the remainder of this section we will assume that $\mathbb{N}^n$ is equipped with some monomial order (the reader may use the degree-lexicographic order), so that $\mathrm{LT}(f)$ is defined and the division algorithm makes sense.

We have seen that the remainder of $f \bmod [g_1, \ldots, g_m]$ obtained from the division algorithm can depend on the order in which the g_i are listed. Informally, a *Gröbner basis* $\{g_1, \ldots, g_m\}$ of the ideal $I = (g_1, \ldots, g_m)$ is a generating set such that, for every m-tuple $G_\sigma = [g_{\sigma(1)}, \ldots, g_{\sigma(m)}]$ formed from the g_i, where $\sigma \in S_m$ is a permutation, the remainder of $f \bmod G_\sigma$ is always the obstruction to whether f lies in I. We define Gröbner bases using a property that is more easily checked, and we then show, in Proposition 6.129, that they are characterized by the more interesting obstruction propery just mentioned.

Definition. A set of polynomials $\{g_1, \ldots, g_m\}$ is a ***Gröbner basis***[18] of the ideal $I = (g_1, \ldots, g_m)$ if, for each nonzero $f \in I$, there is some g_i with $\mathrm{LT}(g_i) \mid \mathrm{LT}(f)$.

Note that a Gröbner basis is a *set* of polynomials, not an m-tuple of polynomials. Example 6.128 shows that

$$\{y^2 + z^2, x^2y + yz, z^3 + xy\}$$

is not a Gröbner basis of the ideal $(y^2 + z^2, x^2y + yz, z^3 + xy)$.

Proposition 6.129. *A set $\{g_1, \ldots, g_m\}$ of polynomials is a Gröbner basis of an ideal $I = (g_1, \ldots, g_m)$ if and only if, for each m-tuple $G_\sigma = [g_{\sigma(1)}, \ldots, g_{\sigma(m)}]$, where $\sigma \in S_m$, every $f \in I$ has remainder* 0 mod G_σ.

Proof. Assume there is some permutation $\sigma \in S_m$ and some $f \in I$ whose remainder mod G_σ is not 0. Among all such polynomials, choose f of minimal Degree. Since $\{g_1, \ldots, g_m\}$ is a Gröbner basis, $\mathrm{LT}(g_i) \mid \mathrm{LT}(f)$ for some i; select the smallest $\sigma(i)$ for

[18]B. Buchberger has written in his article in Buchberger-Winkler, *Gröbner Bases and Applications*, "The early paper of Gröbner in 1954, although not yet containing the essential ingredients of Gröbner basis theory, pointed in the right direction and motivated me, in 1976, to assign the name of W. Gröbner (1899–1980) to the theory."

which there is a reduction $f \overset{g_{\sigma(i)}}{\to} h$, and note that $h \in I$. Since $\text{Deg}(h) \prec \text{Deg}(f)$, by Proposition 6.126, the division algorithm gives a sequence of reductions $h = h_0 \to h_1 \to h_2 \to \cdots \to h_p = 0$. But the division algorithm for f adjoins $f \to h$ at the front, showing that 0 is the remainder of f mod G_σ, a contradiction.

Conversely, assume that every $f \in I$ has remainder 0 mod G_σ but that $\{g_1, \ldots, g_m\}$ is not a Gröbner basis of $I = (g_1, \ldots, g_m)$. If there is a nonzero $f \in I$ with $\text{LT}(g_i) \nmid \text{LT}(f)$ for every i, then in any reduction $f \overset{g_i}{\to} h$, we have $\text{LT}(h) = \text{LT}(f)$. Hence, if $G = [g_1, \ldots, g_m]$, the division algorithm mod G gives reductions $f \to h_1 \to h_2 \to \cdots \to h_p = r$ in which $\text{LT}(r) = \text{LT}(f)$. Therefore, $r \neq 0$; that is, the remainder of f mod G is not zero, and this is a contradiction. •

Corollary 6.130. *If $\{g_1, \ldots, g_m\}$ is a Gröbner basis of the ideal $I = (g_1, \ldots, g_m)$, and if $G = [g_1, \ldots, g_m]$ is any m-tuple formed from the g_i, then for every $f(X) \in k[X]$, there is a unique $r(X) \in k[X]$, which is reduced* mod $\{g_1, \ldots, g_m\}$, *such that $f - r \in I$; in fact, r is the remainder of f* mod G.

Proof. The division algorithm gives a polynomial r, reduced mod $\{g_1, \ldots, g_m\}$, and polynomials $a_1, \ldots, a_m$ with $f = a_1g_1 + \cdots + a_mg_m + r$; clearly, $f - r = a_1g_1 + \cdots + a_mg_m \in I$.

To prove uniqueness, suppose that r and r' are reduced mod $\{g_1, \ldots, g_m\}$ and that $f - r$ and $f - r'$ lie in I, so that $(f - r') - (f - r) = r - r' \in I$. Since r and r' are reduced mod $\{g_1, \ldots, g_m\}$, none of their terms is divisible by any $\text{LT}(g_i)$. If $r - r' \neq 0$, then Exercise 6.82 on page 410 says that no term of $r - r'$ is divisible by any $\text{LT}(g_i)$; in particular, $\text{LT}(r - r')$ is not divisible by any $\text{LT}(g_i)$, and this contradicts Proposition 6.129. Therefore, $r = r'$. •

The next corollary shows that Gröbner bases resolve the problem of different remainders in the division algorithm arising from different m-tuples.

Corollary 6.131. *Let $\{g_1, \ldots, g_m\}$ be a Gröbner basis of the ideal $I = (g_1, \ldots, g_m)$, and let $G = [g_1, \ldots, g_m]$.*

(i) *If $f(X) \in k[X]$ and $G_\sigma = [g_{\sigma(1)}, \ldots, g_{\sigma(m)}]$, where $\sigma \in S_m$ is a permutation, then the remainder of f* mod G *is equal to the remainder of f* mod G_σ.

(ii) *A polynomial $f \in I$ if and only if f has remainder* 0 mod G.

Proof. (i) If r is the remainder of f mod G, then Corollary 6.130 says that r is the unique polynomial, reduced mod $\{g_1, \ldots, g_m\}$, with $f - r \in I$; similarly, the remainder r_σ of f mod G_σ is the unique polynomial, reduced mod $\{g_1, \ldots, g_m\}$, with $f - r_\sigma \in I$. The uniqueness assertion in Corollary 6.130 gives $r = r_\sigma$.

(ii) Proposition 6.129 shows that if $f \in I$, then its remainder is 0. For the converse, if r is the remainder of f mod G, then $f = q + r$, where $q \in I$. Hence, if $r = 0$, then $f \in I$. •

There are several obvious questions. Do Gröbner bases exist and, if they do, are they unique? Given an ideal I in $k[X]$, is there an algorithm to find a Gröbner basis of I?

The notion of *S-polynomial* will allow us to recognize a Gröbner basis, but we first introduce some notation.

Definition. If $\alpha = (\alpha_1, \ldots, \alpha_n)$ and $\beta = (\beta_1, \ldots, \beta_n)$ are in $\mathbb{N}^n$, define

$$\alpha \vee \beta = \mu,$$

where $\mu_i = \max\{\alpha_i, \beta_i\}$ and $\mu = (\mu_1, \ldots, \mu_n)$.

Note that $X^{\alpha \vee \beta}$ is the least common multiple of the monomials X^α and X^β.

Definition. Let $f(X), g(X) \in k[X]$, where $\text{LT}(f) = a_\alpha X^\alpha$ and $\text{LT}(g) = b_\beta X^\beta$. Define

$$L(f, g) = X^{\alpha \vee \beta}.$$

The ***S-polynomial*** $S(f, g)$ is defined by

$$S(f, g) = \frac{L(f, g)}{\text{LT}(f)} f - \frac{L(f, g)}{\text{LT}(g)} g;$$

that is, if $\mu = \alpha \vee \beta$, then

$$S(f, g) = a_\alpha^{-1} X^{\mu - \alpha} f(X) - b_\beta^{-1} X^{\mu - \beta} g(X).$$

Note that $S(f, g) = -S(g, f)$.

Example 6.132.
(i) If $f(x, y) = 3x^2y$ and $g(x, y) = 5xy^3 - y$ (in degree-lexicographic order), then $L(f, g) = x^2y^3$ and

$$S(f, g) = \frac{x^2y^3}{3x^2y} 3x^2y - \frac{x^2y^3}{5xy^3}(5xy^3 - y) = \tfrac{1}{5}xy.$$

(ii) If $f(X)$ and $g(X)$ are monomials, say, $f(X) = a_\alpha X^\alpha$ and $g(X) = b_\beta X^\beta$, then

$$S(f, g) = \frac{X^{\alpha \vee \beta}}{a_\alpha X^\alpha} a_\alpha X^\alpha - \frac{X^{\alpha \vee \beta}}{b_\beta X^\beta} b_\beta X^\beta = 0. \quad ◀$$

The following technical lemma indicates why S-polynomials are relevant. It says that if $\text{Deg}(\sum_j a_j g_j) \prec \delta$, where the a_j are monomials, while $\text{Deg}(a_j g_j) = \delta$ for all j, then any polynomial of multidegree $\prec \delta$ can be rewritten as a linear combination of S-polynomials, with monomial coefficents, each of whose terms has multidegree strictly less than δ.

Lemma 6.133. *Given $g_1(X), \dots, g_\ell(X) \in k[X]$ and monomials $c_j X^{\alpha(j)}$, let $h(X) = \sum_{j=1}^{\ell} c_j X^{\alpha(j)} g_j(X)$.*

Let δ be a multidegree. If $\mathrm{Deg}(h) \prec \delta$ and $\mathrm{Deg}(c_j X^{\alpha(j)} g_j(X)) = \delta$ for all $j \le \ell$, then there are $d_j \in k$ with

$$h(X) = \sum_j d_j X^{\delta - \mu(j)} S(g_j, g_{j+1}),$$

where $\mu(j) = \mathrm{Deg}(g_j) \vee \mathrm{Deg}(g_{j+1})$, and for all $j < \ell$,

$$\mathrm{Deg}\left(X^{\delta-\mu(j)} S(g_j, g_{j+1})\right) \prec \delta.$$

Proof. Let $\mathrm{LT}(g_j) = b_j X^{\beta(j)}$, so that $\mathrm{LT}(c_j X^{\alpha(j)} g_j(X)) = c_j b_j X^\delta$. The coefficient of X^δ in $h(X)$ is thus $\sum_j c_j b_j$. Since $\mathrm{Deg}(h) \prec \delta$, we must have $\sum_j c_j b_j = 0$. Define monic polynomials

$$u_j(X) = b_j^{-1} X^{\alpha(j)} g_j(X).$$

There is a telescoping sum

$$\begin{aligned} h(X) &= \sum_{j=1}^{\ell} c_j X^{\alpha(j)} g_j(X) \\ &= \sum_{j=1}^{\ell} c_j b_j u_j \\ &= c_1 b_1 (u_1 - u_2) + (c_1 b_1 + c_2 b_2)(u_2 - u_3) + \cdots \\ &\qquad + (c_1 b_1 + \cdots + c_{\ell-1} b_{\ell-1})(u_{\ell-1} - u_\ell) \\ &\qquad + (c_1 b_1 + \cdots + c_\ell b_\ell) u_\ell. \end{aligned}$$

The last term $(c_1 b_1 + \cdots + c_\ell b_\ell) u_\ell = 0$, for $\sum_j c_j b_j = 0$. Since $\delta = \mathrm{Deg}(c_j X^{\alpha(j)} g_j(X))$, we have $\alpha(j) + \beta(j) = \delta$, so that $X^{\beta(j)} \mid X^\delta$ for all j. Hence, for all $j < \ell$, we have $\mathrm{lcm}\{X^{\beta(j)}, X^{\beta(j+1)}\} = X^{\beta(j) \vee \beta(j+1)} \mid X^\delta$; that is, if we write $\mu(j) = \beta(j) \vee \beta(j+1)$, then $\delta - \mu(j) \in \mathbb{N}^n$. But

$$\begin{aligned} X^{\delta-\mu(j)} S(g_j, g_{j+1}) &= X^{\delta-\mu(j)} \Big(\frac{X^{\mu(j)}}{\mathrm{LT}(g_j)} g_j(X) - \frac{X^{\mu(j)}}{\mathrm{LT}(g_{j+1})} g_{j+1}(X) \Big) \\ &= \frac{X^\delta}{\mathrm{LT}(g_j)} g_j(X) - \frac{X^\delta}{\mathrm{LT}(g_{j+1})} g_{j+1}(X) \\ &= b_j^{-1} X^{\alpha(j)} g_j - b_{j+1}^{-1} X^{\alpha(j+1)} g_{j+1} \\ &= u_j - u_{j+1}. \end{aligned}$$

Substituting this equation into the telescoping sum gives a sum of the desired form, where $d_j = c_1 b_1 + \cdots + c_j b_j$:

$$\begin{aligned} h(X) = c_1 b_1 X^{\delta-\mu(1)} S(g_1, g_2) &+ (c_1 b_1 + c_2 b_2) X^{\delta-\mu(2)} S(g_2, g_3) + \cdots \\ &+ (c_1 b_1 + \cdots + c_{\ell-1} b_{\ell-1}) X^{\delta-\mu(\ell-1)} S(g_{\ell-1}, g_\ell). \end{aligned}$$

Finally, since both u_j and u_{j+1} are monic with leading term of multidegree δ, we have $\mathrm{Deg}(u_j - u_{j+1}) \prec \delta$. But we have shown that $u_j - u_{j+1} = X^{\delta-\mu(j)}S(g_j, g_{j+1})$, and so $\mathrm{Deg}(X^{\delta-\mu(j)}S(g_j, g_{j+1})) \prec \delta$, as desired. •

By Proposition 6.129, $\{g_1, \ldots, g_m\}$ is a Gröbner basis of $I = (g_1, \ldots, g_m)$ if every $f \in I$ has remainder 0 mod G (where G is any m-tuple formed by ordering the g_i). The importance of the next theorem lies in its showing that it is necessary to compute the remainders of only finitely many polynomials, namely, the S-polynomials, to determine whether $\{g_1, \ldots, g_m\}$ is a Gröbner basis.

Theorem 6.134 (Buchberger). *A set $\{g_1, \ldots, g_m\}$ is a Gröbner basis of an ideal $I = (g_1, \ldots, g_m)$ if and only if $S(g_p, g_q)$ has remainder* 0 mod G *for all p, q, where $G = [g_1, \ldots, g_m]$.*

Proof. Clearly, $S(g_p, g_q)$, being a linear combination of g_p and g_q, lies in I. Hence, if $G = \{g_1, \ldots, g_m\}$ is a Gröbner basis, then $S(g_p, g_q)$ has remainder 0 mod G, by Proposition 6.129.

Conversely, assume that $S(g_p, g_q)$ has remainder 0 mod G for all p, q; we must show that every $f \in I$ has remainder 0 mod G. By Proposition 6.129, it suffices to show that if $f \in I$, then $\mathrm{LT}(g_i) \mid \mathrm{LT}(f)$ for some i. Since $f \in I = (g_1, \ldots, g_m)$, we may write $f = \sum_i h_i g_i$, and so

$$\mathrm{Deg}(f) \preceq \max_i\{\mathrm{Deg}(h_i g_i)\}.$$

If there is equality, then $\mathrm{Deg}(f) = \mathrm{Deg}(h_i g_i)$ for some i, and so Proposition 6.125 gives $\mathrm{LT}(g_i) \mid \mathrm{LT}(f)$, as desired. Therefore, we may assume strict inequality: $\mathrm{Deg}(f) \prec \max_i\{\mathrm{Deg}(h_i g_i)\}$.

The polynomial f may be written as a linear combination of the g_i in many ways. Of all the expressions of the form $f = \sum_i h_i g_i$, choose one in which $\delta = \max_i\{\mathrm{Deg}(h_i g_i)\}$ is minimal (which is possible because $\preceq$ is a well-order). If $\mathrm{Deg}(f) = \delta$, we are done, as we have seen; therefore, we may assume that there is strict inequality: $\mathrm{Deg}(f) \prec \delta$. Write

$$f = \sum_{\substack{j \\ \mathrm{Deg}(h_j g_j) = \delta}} h_j g_j + \sum_{\substack{\ell \\ \mathrm{Deg}(h_\ell g_\ell) \prec \delta}} h_\ell g_\ell. \tag{7}$$

If $\mathrm{Deg}(\sum_j h_j g_j) = \delta$, then $\mathrm{Deg}(f) = \delta$, a contradiction; hence, $\mathrm{Deg}(\sum_j h_j g_j) \prec \delta$. But the coefficient of X^δ in this sum is obtained from its leading terms, so that

$$\mathrm{Deg}\Big(\sum_j \mathrm{LT}(h_j) g_j\Big) \prec \delta.$$

Now $\sum_j \mathrm{LT}(h_j) g_j$ is a polynomial satisfying the hypotheses of Lemma 6.133, and so there are constants d_j and multidegrees $\mu(j)$ so that

$$\sum_j \mathrm{LT}(h_j) g_j = \sum_j d_j X^{\delta-\mu(j)} S(g_j, g_{j+1}), \tag{8}$$

where $\operatorname{Deg}\left(X^{\delta-\mu(j)}S(g_j, g_{j+1})\right) \prec \delta$.[19]

Since each $S(g_j, g_{j+1})$ has remainder 0 mod G, the division algorithm gives $a_{ji}(X) \in k[X]$ with $S(g_j, g_{j+1}) = \sum_i a_{ji}g_i$, where $\operatorname{Deg}(a_{ji}g_i) \preceq \operatorname{Deg}(S(g_j, g_{j+1}))$ for all j, i. It follows that

$$X^{\delta-\mu(j)}S(g_j, g_{j+1}) = \sum_i X^{\delta-\mu(j)}a_{ji}g_i.$$

Therefore, Lemma 6.133 gives

$$\operatorname{Deg}(X^{\delta-\mu(j)}a_{ji}g_i) \preceq \operatorname{Deg}(X^{\delta-\mu(j)}S(g_j, g_{j+1})) \prec \delta. \tag{9}$$

Substituting into Eq. (8), we have

$$\begin{aligned}\sum_j \mathrm{LT}(h_j)g_j &= \sum_j d_j X^{\delta-\mu(j)}S(g_j, g_{j+1})\\ &= \sum_j d_j\Big(\sum_i X^{\delta-\mu(j)}a_{ji}g_i\Big)\\ &= \sum_i\Big(\sum_j d_j X^{\delta-\mu(j)}a_{ji}\Big)g_i.\end{aligned}$$

If we denote $\sum_j d_j X^{\delta-\mu(j)}a_{ji}$ by h_i', then

$$\sum_j \mathrm{LT}(h_j)g_j = \sum_i h_i' g_i, \tag{10}$$

where, by Eq. (9), $\operatorname{Deg}(h_i' g_i) \prec \delta$ for all i.

Finally, we substitute the expression in Eq. (10) into Eq. (7):

$$\begin{aligned} f &= \sum_{\substack{j\\ \operatorname{Deg}(h_j g_j)=\delta}} h_j g_j + \sum_{\substack{\ell\\ \operatorname{Deg}(h_\ell g_\ell)\prec\delta}} h_\ell g_\ell\\ &= \sum_{\substack{j\\ \operatorname{Deg}(h_j g_j)=\delta}} \mathrm{LT}(h_j)g_j + \sum_{\substack{j\\ \operatorname{Deg}(h_j g_j)=\delta}} [h_j - \mathrm{LT}(h_j)]g_j + \sum_{\substack{\ell\\ \operatorname{Deg}(h_\ell g_\ell)\prec\delta}} h_\ell g_\ell\\ &= \sum_i h_i' g_i + \sum_{\substack{j\\ \operatorname{Deg}(h_j g_j)=\delta}} [h_j - \mathrm{LT}(h_j)]g_j + \sum_{\substack{\ell\\ \operatorname{Deg}(h_\ell g_\ell)\prec\delta}} h_\ell g_\ell.\end{aligned}$$

We have rewritten f as a linear combination of the g_i in which each term has multidegree strictly smaller than δ, contradicting the minimality of δ. This completes the proof. •

---

[19]The reader may wonder why we consider all S-polynomials $S(g_p, g_q)$ instead of only those of the form $S(g_i, g_{i+1})$. The answer is that the remainder condition is applied only to those $h_j g_j$ for which $\operatorname{Deg}(h_j g_j) = \delta$, and so the indices viewed as i's need not be consecutive.

Corollary 6.135. *If $I = (f_1, \ldots, f_s)$ in $k[X]$, where each f_i is a monomial (that is, if I is a monomial ideal), then $\{f_1, \ldots, f_s\}$ is a Gröbner basis of I.*

Proof. By Example 6.132(ii), the S-polynomial of any pair of monomials is 0. •

Here is the main result: A Gröbner basis of $(f_1, \ldots, f_s)$ can obtained by adjoining remainders of S-polynomials.

Theorem 6.136 (Buchberger's Algorithm). *Every ideal $I = (f_1, \ldots, f_s)$ in $k[X]$ has a Gröbner basis*[20] *that can be computed by an algorithm.*

Proof. Here is a pseudocode for an algorithm.

```
Input : B = { f_1, ..., f_s }    G = [ f_1, ..., f_s ]
Output : a Gröbner basis B = {g_1, ..., g_m}  containing { f_1, ..., f_s }
B := { f_1, ..., f_s }    G := [ f_1, ..., f_s ]
REPEAT
        B' := B    G' := G
        FOR each  pair g, g' with g ≠ g' ∈ B' DO
                r :=  remainder  of S(g, g')  mod G'
                IF r ≠ 0
                   THEN B := B ∪ {r} and G' = [g_1, ..., g_m, r]
UNTIL B = B'
```

Now each loop of the algorithm enlarges a subset $B \subseteq I = (g_1, \ldots, g_m)$ by adjoining the remainder mod G of one of its S-polynomials $S(g, g')$. As $g, g' \in I$, the remainder r of $S(g, g')$ lies in I, and so the larger set $B \cup \{r\}$ is contained in I.

The only obstruction to the algorithm's stopping at some B' is if some $S(g, g')$ does not have remainder 0 mod G'. Thus, if the algorithm stops, then Theorem 6.134 shows that B' is a Gröbner basis.

To see that the algorithm does stop, suppose a loop starts with B' and ends with B. Since $B' \subseteq B$, we have an inclusion of monomial ideals

$$\big(\mathrm{LT}(g') \colon g' \in B'\big) \subseteq (\mathrm{LT}(g) \colon g \in B).$$

We claim that if $B' \subsetneq B$, then there is also a strict inclusion of ideals. Suppose that r is a (nonzero) remainder of some S-polynomial mod B', and that $B = B' \cup \{r\}$. By definition, the remainder r is reduced mod G', and so no term of r is divisible by $\mathrm{LT}(g')$ for any $g' \in B'$; in particular, $\mathrm{LT}(r)$ is not divisible by any $\mathrm{LT}(g')$. Hence, $\mathrm{LT}(r) \notin (\mathrm{LT}(g') \colon g' \in B')$, by Exercise 6.83 on page 410. On the other hand, we do have $\mathrm{LT}(r) \in (\mathrm{LT}(g) \colon g \in B)$. Therefore, if the algorithm does not stop, there is an infinite strictly ascending chain of ideals in $k[X]$, and this contradicts the Hilbert basis theorem, for $k[X]$ has the ACC. •

[20] A nonconstructive proof of the existence of a Gröbner basis can be given using the proof of the Hilbert basis theorem; for example, see Section 2.5 of Cox–Little–O'Shea, *Ideals, Varieties, and Algorithms* (they also give a constructive proof in Section 2.7).

Example 6.137.
The reader may show that $B' = \{y^2 + z^2, x^2y + yz, z^3 + xy\}$ is not a Gröbner basis because $S(y^2+z^2, x^2y+yz) = x^2z^2 - y^2z$ does not have remainder 0 mod G'. However, adjoining $x^2z^2 - y^2z$ does give a Gröbner basis B because all the S-polynomials in B [there are $\binom{4}{2} = 6$ of them] have remainder 0 mod B'. ◄

Theoretically, Buchberger's algorithm computes a Gröbner basis, but the question arises how practical it is. In very many cases, it does compute in a reasonable amount of time; on the other hand, there are examples in which it takes a very long time to produce its output. The efficiency of Buchberger's algorithm is discussed in Section 2.9 of Cox–Little–O'Shea, *Ideals, Varieties, and Algorithms*.

Corollary 6.138.

(i) *If $I = (f_1, \ldots, f_t)$ is an ideal in $k[X]$, then there is an algorithm to determine whether a polynomial $h(X) \in k[X]$ lies in I.*

(ii) *If $I = (f_1, \ldots, f_t) \subseteq k[X]$, then there is an algorithm to determine whether a polynomial $g(X) \in k[X]$ lies in $\sqrt{I}$.*

(iii) *If $I = (f_1, \ldots, f_t)$ and $I' = (f_1', \ldots, f_s')$ are ideals in $k[X]$, then there is an algorithm to determine whether $I = I'$.*

Proof. (i) Use Buchberger's algorithm to find a Gröbner basis B of I, and then use the division algorithm to compute the remainder of h mod G (where G is any m-tuple arising from ordering the polynomials in B). By Corollary 6.131(ii), $h \in I$ if and only if $r = 0$.

(ii) Use Exercise 6.66 on page 398 and then use Buchberger's algorithm to find a Gröbner basis of $(f_1, \ldots, f_t, 1 - yg)$ in $k[X, y]$.

(iii) Use Buchberger's algorithm to find Gröbner bases $\{g_1, \ldots, g_m\}$ and $\{g_1', \ldots, g_m'\}$ of I and I', respectively. By part (i), there is an algorithm to determine whether each $g_j' \in I$, and $I' \subseteq I$ if each $g_j' \in I$. Similarly, there is an algorithm to determine the reverse inclusion, and so there is an algorithm to determine whether $I = I'$. •

A Gröbner basis $B = \{g_1, \ldots, g_m\}$ can be too large. For example, it follows from the very definition of Gröbner basis that if $f \in I$, then $B \cup \{f\}$ is also a Gröbner basis of I; thus, we may seek Gröbner bases that are, in some sense, minimal.

Definition. A basis $\{g_1, \ldots, g_m\}$ of an ideal I is ***reduced*** if

(i) each g_i is monic;

(ii) each g_i is reduced mod $\{g_1, \ldots, \widehat{g_i}, \ldots, g_m\}$.

Exercise 6.90 on page 421 gives an algorithm for computing a reduced basis for every ideal $(f_1, \ldots, f_t)$. When combined with the algorithm in Exercise 6.93 on page 422,

it shrinks a Gröbner basis to a *reduced* Gröbner basis. It can be proved that a reduced Gröbner basis of an ideal is unique. In the special case when each $f_i(X)$ is linear, that is,

$$f_i(X) = a_{i1}x_1 + \cdots + a_{in}x_n.$$

then the common zeros $\mathrm{Var}(f_1, \ldots, f_t)$ are the solutions of a homogeneous system of t equations in n unknowns. If $A = [a_{ij}]$ is the $t \times n$ matrix of coefficients, then it can be shown that the reduced Gröbner basis corresponds to the row-reduced echelon form for the matrix A (see Section 10.5 of Becker–Weispfenning, *Gröbner Bases*). Another special case occurs when when $f_1, \ldots, f_t$ are polynomials in one variable. The reduced Gröbner basis obtained from $\{f_1, \ldots, f_t\}$ turns out to be their gcd, and so the euclidean algorithm has been generalized to polynomials in several variables.

Corollary 6.138 does not begin by saying "If I is an ideal in $k[X]$"; instead, it specifies a basis: $I = (f_1, \ldots, f_t)$. The reason, of course, is that Buchberger's algorithm requires a basis as input. For example, if $J = (h_1, \ldots, h_s)$, then the algorithm cannot be used directly to check whether a polynomial $f(X)$ lies in the radical $\sqrt{J}$, for we do not have a basis of $\sqrt{J}$. The book of Becker–Weispfenning, *Gröbner Bases*, gives an algorithm computing a basis of $\sqrt{J}$ (page 393) when k satisfies certain conditions. There is no algorithm known that computes the associated primes of an ideal, although there are algorithms to do some special cases of this general problem. As we mentioned at the beginning of this section, if an ideal I has a primary decomposition $I = Q_1 \cap \cdots \cap Q_r$, then the associated prime P_i has the form $\sqrt{(I : c_i)}$ for any $c_i \in \bigcap_{j \neq i} Q_j$ and $c_i \notin Q_i$. There is an algorithm computing a basis of colon ideals (Becker–Weispfenning, *Gröbner Bases*, page 266). Thus, we could compute P_i if there were an algorithm finding elements c_i. For a survey of applications of Gröbner bases to various parts of mathematics, the reader should see Buchberger–Winkler, *Gröbner Bases and Applications*.

We end this chapter by showing how to find a basis of an intersection of ideals.

Given a system of polynomial equations in several variables, one way to find solutions is to eliminate variables (van der Waerden, *Modern Algebra* II, Chapter XI). Given an ideal $I \subseteq k[X]$, we are led to an ideal in a subset of the indeterminates, which is essentially the intersection of $\mathrm{Var}(I)$ with a lower-dimensional plane.

Definition. Let k be a field and let $I \subseteq k[X, Y]$ be an ideal, where $k[X, Y]$ is the polynomial ring in disjoint sets of variables $X \cup Y$. The ***elimination ideal*** is

$$I_X = I \cap k[X].$$

For example, if $I = (x^2, xy)$, then a Gröbner basis is $\{x^2, xy\}$ (they are monomials, so that Corollary 6.135 applies), and $I_x = (x^2) \subseteq k[x]$, while $I_y = \{0\}$.

Proposition 6.139. *Let k be a field and let $k[X] = k[x_1, \ldots, x_n]$ have a monomial order for which $x_1 \succ x_2 \succ \cdots \succ x_n$ (for example, the lexicographic order) and, for fixed $p > 1$, let $Y = x_p, \ldots, x_n$. If $I \subseteq k[X]$ has a Gröbner basis $G = \{g_1, \ldots, g_m\}$, then $G \cap I_Y$ is a Gröbner basis for the elimination ideal $I_Y = I \cap k[x_p, \ldots, x_n]$.*

Proof. Recall that $\{g_1, \ldots, g_m\}$ being a Gröbner basis of $I = (g_1, \ldots, g_m)$ means that for each nonzero $f \in I$, there is g_i with $\mathrm{LT}(g_i) \mid \mathrm{LT}(f)$. Let $f(x_p, \ldots, x_n) \in I_Y$ be nonzero. Since $I_Y \subseteq I$, there is some $g_i(X)$ with $\mathrm{LT}(g_i) \mid \mathrm{LT}(f)$; hence, $\mathrm{LT}(g_i)$ involves only the "later" variables $x_p, \ldots, x_n$. Let $\mathrm{Deg}(\mathrm{LT}(g_i)) = \beta$. If g_i has a term $c_\alpha X^\alpha$ involving "early" variables x_i with $i < p$, then $\alpha \succ \beta$, because $x_1 \succ \cdots \succ x_p \succ \cdots \succ x_n$. This is a contradiction, for β, the Degree of the leading term of g_i, is greater than the Degree of any other term of g_i. It follows that $g_i \in k[x_p, \ldots, x_n]$. Exercise 6.92 on page 422 now shows that $G \cap k[x_p, \ldots, x_n]$ is a Gröbner basis for $I_Y = I \cap k[x_p, \ldots, x_n]$. •

We can now give Gröbner bases of intersections of ideals.

Proposition 6.140. *Let k be a field, and let $I_1, \ldots, I_t$ be ideals in $k[X]$, where $X = x_1, \ldots, x_n$.*

(i) *Consider the polynomial ring $k[X, y_1, \ldots, y_t]$ having a new variable y_j for each j with $1 \le j \le t$. If J is the ideal in $k[X, y_1, \ldots, y_t]$ generated by $1 - (y_1 + \cdots + y_t)$ and $y_j I_j$, for all j, then $\bigcap_{j=1}^t I_j = J_X$.*

(ii) *Given Gröbner bases of $I_1, \ldots, I_t$, a Gröbner basis of $\bigcap_{j=1}^t I_j$ can be computed.*

Proof. (i) If $f = f(X) \in J_X = J \cap k[X]$, then $f \in J$, and so there is an equation

$$f(X) = g(X, Y)(1 - \sum y_j) + \sum_j h_j(X, y_1, \ldots, y_t) y_j q_j(X),$$

where $g, h_j \in k[X, Y]$ and $q_j \in I_j$. Setting $y_j = 1$ and the other y's equal to 0 gives $f = h_j(X, 0, \ldots, 1, \ldots, 0) q_j(X)$. Note that $h_j(X, 0, \ldots, 1, \ldots, 0) \in k[X]$, and so $f \in I_j$. As j was arbitrary, we have $f \in \bigcap I_j$, and so $J_X \subseteq \bigcap I_j$.

For the reverse inclusion, if $f \in \bigcap I_j$, then the equation

$$f = f(1 - \sum y_j) + \sum_j y_j f$$

shows that $f \in J_X$, as desired.

(ii) This follows from part (i) and Proposition 6.139 if we use a monomial order in which all the variables in X precede the variables in Y. •

Example 6.141.
Consider the ideal $I = (x) \cap (x^2, xy, y^2) \subseteq k[x, y]$, where k is a field, that we considered in Example 6.117(ii). Even though it is not difficult to find a basis of I by hand, we shall use Gröbner bases to illustrate Proposition 6.140. Let u and v be new variables, and define

$$J = (1 - u - v, ux, vx^2, vxy, vy^2) \subseteq k[x, y, u, v].$$

The first step is to find a Gröbner basis of J; we use the lex monomial order with $x \prec y \prec u \prec v$. Since the S-polynomial of two monomials is 0, Buchberger's algorithm quickly

gives a Gröbner basis[21] G of J:

$$G = \{v + u - 1, x^2, yx, ux, uy^2 - y^2\}.$$

It follows from Proposition 6.139 that a Gröbner basis of I is $G \cap k[x, y]$: all those elements of G that do not involve the variables u and v. Thus,

$$I = (x) \cap (x^2, xy, y^2) = (x^2, xy). \quad \blacktriangleleft$$

We mention that Gröbner bases can be adapted to noncommutative rings. A. I. Shirsov began investigating whether there are analogs holding for rings of polynomials in several noncommuting variables, with the aim of implementing algorithms to solve problems in Lie algebras.

EXERCISES

Use the degree-lexicographic monomial order in the following exercises.

6.85 Let $I = (y - x^2, z - x^3)$.

(i) Order $x \prec y \prec z$, and let $\preceq_{\text{lex}}$ be the corresponding monomial order on $\mathbb{N}^3$. Prove that $[y - x^2, z - x^3]$ is not a Gröbner basis of I.

(ii) Order $y \prec z \prec x$, and let $\preceq_{\text{lex}}$ be the corresponding monomial order on $\mathbb{N}^3$. Prove that $[y - x^2, z - x^3]$ is a Gröbner basis of I.

6.86 Find a Gröbner basis of $I = (x^2 - 1, xy^2 - x)$.

6.87 Find a Gröbner basis of $I = (x^2 + y, x^4 + 2x^2y + y^2 + 3)$.

6.88 Find a Gröbner basis of $I = (xz, xy - z, yz - x)$. Does $x^3 + x + 1$ lie in I?

6.89 Find a Gröbner basis of $I = (x^2 - y, y^2 - x, x^2y^2 - xy)$. Does $x^4 + x + 1$ lie in I?

6.90 Show that the following pseudocode gives a reduced basis Q of an ideal $I = (f_1, \ldots, f_t)$.

```
Input: P = [f_1, ..., f_t]
Output: Q = [q_1, ..., q_s]
Q := P
WHILE there is q ∈ Q which is
        not reduced mod Q − {q} DO
  select q ∈ Q which is not reduced mod Q − {q}
  Q := Q − {q}
  h := the remainder of q mod Q
  IF h ≠ 0 THEN
    Q := Q ∪ {h}
  END IF
END WHILE
make all q ∈ Q monic
```

[21] This is actually the reduced Gröbner basis given by Exercise 6.93 on page 422.

6.91 If G is a Gröbner basis of an ideal I, and if Q is the basis of I obtained from the algorithm in Exercise 6.90, prove that Q is also a Gröbner basis of I.

6.92 Let I be an ideal in $k[X]$, where k is a field and $k[X]$ has a monomial order. Prove that if a set of polynomials $\{g_1, \ldots, g_m\} \subseteq I$ has the property that, for each nonzero $f \in I$, there is some g_i with $\mathrm{LT}(g_i) \mid \mathrm{LT}(f)$, then $I = (g_1, \ldots, g_m)$. Conclude, in the definition of Gröbner basis, that one need not assume that I is generated by $g_1, \ldots, g_m$.

6.93 Show that the following pseudocode replaces a Gröbner basis G with a reduced Gröbner basis H.

```
Input: G = {g_1, ..., g_m}
Output: H
H := ∅;   F := G
WHILE F ≠ ∅ DO
  select f' from F
  F := F − {f'}
  IF LT(f) ∤ LT(f') for all f ∈ F AND
        LT(h) ∤ LT(f') for all h ∈ H THEN
    H := H ∪ {f'}
  END IF
END WHILE
apply the algorithm in Exercise 6.90 to H
```

7

Modules and Categories

We now introduce *R-modules*, where R is a commutative ring; formally, they generalize vector spaces in the sense that scalars are allowed to be in R instead of a field. If R is a PID, then we shall see, in Chapter 9, that classification of finitely generated R-modules simultaneously gives a classification of all finitely generated abelian groups as well as a classification of all linear transformations on a finite-dimensional vector space by canonical forms. After introducing noncommutative rings in Chapter 8, we will define modules over these rings, and they will be used, in an essential way, to prove that every finite group of order $p^m q^n$, where p and q are primes, is a solvable group.

Categories and functors first arose in algebraic topology, where topological spaces and continuous maps are studied by means of certain algebraic systems (homology groups, cohomology rings, homotopy groups) associated to them. Categorical notions have proven to be valuable in purely algebraic contexts as well; indeed, it is fair to say that the recent great strides in algebraic geometry could not have occurred outside a categorical setting.

7.1 Modules

An R-module is just a "vector space over a ring R"; that is, in the definition of vector space, allow the scalars to be in R instead of in a field.

Definition. Let R be a commutative ring. An ***R-module*** is an (additive) abelian group M equipped with a ***scalar multiplication*** $R \times M \to M$, denoted by

$$(r, m) \mapsto rm,$$

such that the following axioms hold for all $m, m' \in M$ and all $r, r', 1 \in R$:

(i) $r(m + m') = rm + rm'$;

(ii) $(r + r')m = rm + r'm$;

(iii) $(rr')m = r(r'm)$;

(iv) $1m = m$.

Remark. This definition also makes sense for noncommutative rings R, in which case M is called a ***left R-module***. ◂

Example 7.1.
(i) Every vector space over a field k is a k-module.

(ii) By the laws of exponents, Proposition 2.23, every abelian group is a $\mathbb{Z}$-module.

(iii) Every commutative ring R is a module over itself if we define scalar multiplication $R \times R \to R$ to be the given multiplication of elements of R. More generally, every ideal I in R is an R-module, for if $i \in I$ and $r \in R$, then $ri \in I$.

(iv) If S is a subring of a commutative ring R, then R is an S-module, where scalar multiplication $S \times R \to R$ is just the given multiplication $(s, r) \mapsto sr$. For example, if k is a commutative ring, then $k[X]$ is a k-module.

(v) Let $T : V \to V$ be a linear transformation, where V is a finite-dimensional vector space over a field k. The vector space V can be made into a $k[x]$-module if scalar multiplication $k[x] \times V \to V$ is defined as follows: If $f(x) = \sum_{i=0}^{m} c_i x^i$ lies in $k[x]$, then

$$f(x)v = \Big(\sum_{i=0}^{m} c_i x^i\Big)v = \sum_{i=0}^{m} c_i T^i(v),$$

where T^0 is the identity map 1_V, and T^i is the composite of T with itself i times if $i \geq 1$. We denote V viewed as a $k[x]$-module by V^T.

Here is a special case of this construction. Let A be an $n \times n$ matrix with entries in k, and let $T : k^n \to k^n$ be the linear transformation $T(w) = Aw$, where w is an $n \times 1$ column vector and Aw is matrix multiplication. Now the vector space k^n is a $k[x]$-module if we define scalar multiplication $k[x] \times k^n \to k^n$ as follows: If $f(x) = \sum_{i=0}^{m} c_i x^i \in k[x]$, then

$$f(x)w = \Big(\sum_{i=0}^{m} c_i x^i\Big)w = \sum_{i=0}^{m} c_i A^i w,$$

where $A^0 = I$ is the identity matrix, and A^i is the ith power of A if $i \geq 1$. We now show that $(k^n)^T = (k^n)^A$. Both modules are comprised of the same elements (namely, all n-tuples), and the scalar multiplications coincide: In $(k^n)^T$, we have $xw = T(w)$; in $(k^n)^A$, we have $xw = Aw$; these are the same because $T(w) = Aw$. ◂

Here is the appropriate notion of homomorphism.

Definition. If R is a ring and M and N are R-modules, then a function $f : M \to N$ is an ***R-homomorphism*** (or ***R-map***) if, for all $m, m' \in M$ and all $r \in R$,

(i) $f(m + m') = f(m) + f(m')$;

(ii) $f(rm) = rf(m)$.

If an R-homomorphism is a bijection, then it is called an ***R-isomorphism***; R-modules M and N are called ***isomorphic***, denoted by $M \cong N$, if there is some R-isomorphism $f\colon M \to N$.

Note that the composite of R-homomorphisms is an R-homomorphism and, if f is an R-isomorphism, then its inverse function f^{-1} is also an R-isomorphism.

Example 7.2.
(i) If R is a field, then R-modules are vector spaces and R-maps are linear transformations. Isomorphisms here are nonsingular linear transformations.

(ii) By Example 7.1(ii), $\mathbb{Z}$-modules are just abelian groups, and Lemma 2.52 shows that every homomorphism of (abelian) groups is a $\mathbb{Z}$-map.

(iii) If M is an R-module and $r \in R$, then ***multiplication by r*** (or *homothety by* r) is the function $\mu_r\colon M \to M$ given by $m \mapsto rm$.

The functions μ_r are R-maps because R is commutative: If $a \in R$ and $m \in M$, then $\mu_r(am) = ram$ while $a\mu_r(m) = arm$.

(iv) Let $T\colon V \to V$ be a linear transformation on a vector space V over a field k, let $v_1, \ldots, v_n$ be a basis of V, and let A be the matrix of T relative to this basis. We now show that the two $k[x]$-modules V^T and $(k^n)^A$ are isomorphic.

Define $\varphi\colon V \to k^n$ by $\varphi(v_i) = e_i$, where $e_1, \ldots, e_n$ is the standard basis of k^n; the linear transformation φ is an isomorphism of vector spaces. To see that φ is a $k[x]$-map, it suffices to prove that $\varphi(f(x)v) = f(x)\varphi(v)$ for all $f(x) \in k[x]$ and all $v \in V$. Now

$$\begin{aligned}\varphi(xv_i) &= \varphi(T(v_i))\\ &= \varphi\Big(\sum a_{ji}v_j\Big)\\ &= \sum a_{ji}\varphi(v_j)\\ &= \sum a_{ji}e_j,\end{aligned}$$

which is the ith column of A. On the other hand,

$$x\varphi(v_i) = A\varphi(v_i) = Ae_i,$$

which is also the ith column of A. It follows that $\varphi(xv) = x\varphi(v)$ for all $v \in V$, and we can easily prove, by induction on $\deg(f)$, that $\varphi(f(x)v) = f(x)\varphi(v)$ for all $f(x) \in k[x]$ and all $v \in V$. ◄

The next proposition generalizes the last example.

Proposition 7.3. *Let V be a vector space over a field k, and let $T, S\colon V \to V$ be linear transformations. Then the $k[x]$-modules V^T and V^S in* Example 7.1(v) *are $k[x]$-isomorphic if and only if there is a vector space isomorphism $\varphi\colon V \to V$ with*

$$S = \varphi T \varphi^{-1}.$$

Proof. If $\varphi\colon V^T \to V^S$ is a $k[x]$-isomorphism, then $\varphi\colon V \to V$ is an isomorphism of vector spaces with

$$\varphi(f(x)v) = f(x)\varphi(v)$$

for all $v \in V$ and all $f(x) \in k[x]$. In particular, if $f(x) = x$, then

$$\varphi(xv) = x\varphi(v).$$

But the definition of scalar multiplication in V^T is $xv = T(v)$, while the definition of scalar multiplication in V^S is $xv = S(v)$. Hence, for all $v \in V$, we have

$$\varphi(T(v)) = S(\varphi(v)).$$

Therefore,

$$\varphi T = S\varphi.$$

As φ is an isomorphism, we have the desired equation $S = \varphi T\varphi^{-1}$.

Conversely, we may assume $\varphi(f(x)v) = f(x)\varphi(v)$ in the special cases $\deg(f) \le 1$:

$$\varphi(xv) = \varphi T(v) = S\varphi(v) = x\varphi(v).$$

Next, an easy induction shows that $\varphi(x^n v) = x^n\varphi(v)$, and a second easy induction, on $\deg(f)$, shows that $\varphi(f(x)v) = f(x)\varphi(v)$. •

It is worthwhile making a special case of the proposition explicit. The next corollary shows how comfortably similarity of matrices fits into the language of modules (and we will see, in Chapter 9, how this contributes to finding canonical forms).

Corollary 7.4. *Let k be a field, and let A and B be $n \times n$ matrices with entries in k. Then the $k[x]$-modules $(k^n)^A$ and $(k^n)^B$ in* Example 7.1(v) *are $k[x]$-isomorphic if and only if there is a nonsingular matrix P with*

$$B = PAP^{-1}.$$

Proof. Define $T : k^n \to k^n$ by $T(y) = Ay$, where $y \in k^n$ is a column; by Example 7.1(v), the $k[x]$-module $(k^n)^T = (k^n)^A$. Similarly, define $S\colon k^n \to k^n$ by $S(y) = By$, and denote the corresponding $k[x]$-module by $(k^n)^B$. The proposition now gives an isomorphism $\varphi\colon V^T \to V^S$ with

$$\varphi(Ay) = B\varphi(y).$$

By Proposition 3.94, there is an $n \times n$ matrix P with $\varphi(y) = Py$ for all $y \in k^n$ (which is nonsingular because φ is an isomorphism). Therefore,

$$PAy = BPy$$

for all $y \in k^n$, and so

$$PA = BP;$$

hence, $B = PAP^{-1}$.

Conversely, the nonsingular matrix P gives an isomorphism $\varphi\colon k^n \to k^n$ by $\varphi(y) = Py$ for all $y \in k^n$. The proposition now shows that $\varphi\colon (k^n)^A \to (k^n)^B$ is a $k[x]$-module isomorphism. •

Homomorphisms can be added.

Definition. If M and N are R-modules, then

$$\mathrm{Hom}_R(M, N) = \{\text{all } R\text{-homomorphisms } M \to N\}.$$

If $f, g \in \mathrm{Hom}_R(M, N)$, then define $f + g \colon M \to N$ by

$$f + g \colon m \mapsto f(m) + g(m).$$

Proposition 7.5. *If M and N are R-modules, where R is a commutative ring, then $\mathrm{Hom}_R(M, N)$ is an R-module, where addition has just been defined, and scalar multiplication is given by*

$$rf \colon m \mapsto f(rm).$$

Moreover, there are distributive laws: If $p \colon M' \to M$ and $q \colon N \to N'$, then

$$(f + g)p = fp + gp \quad \text{and} \quad q(f + g) = qf + qg$$

for all $f, g \in \mathrm{Hom}_R(M, N)$.

Proof. Verification of the axioms in the definition of R-module is straightforward, but we present the proof of

$$(rr')f = r(r'f)$$

because it uses commutativity of R.

If $m \in M$, then $(rr')f \colon m \mapsto f(rr'm)$. On the other hand, $r(r'f) \colon m \mapsto (r'f)(rm) = f(r'rm)$. Since R is commutative, $rr' = r'r$, and so $(rr')f = r(r'f)$. •

Example 7.6.

In linear algebra, a ***linear functional*** on a vector space V over a field k is a linear transformation $\varphi \colon V \to k$ [after all, k is a (one-dimensional) vector space over itself]. For example, if

$$V = \{\text{continuous } f \colon [0, 1] \to \mathbb{R}\},$$

then integration, $f \mapsto \int_0^1 f(t)\,dt$, is a linear functional on V.

If V is a vector space over a field k, then its ***dual space*** is the set of all linear functionals on V:

$$V^* = \mathrm{Hom}_k(V, k).$$

By the proposition, V^* is also a k-module; that is, V^* is a vector space over k. ◂

We now show that constructions made for abelian groups and for vector spaces can also be made for modules. A *submodule* S is an R-module contained in a larger R-module M such that if $s, s' \in S$ and $r \in R$, then $s + s'$ and rs have the same meaning in S as in M.

Definition. If M is an R-module, then a ***submodule*** N of M, denoted by $N \subseteq M$, is an additive subgroup N of M closed under scalar multiplication: $rn \in N$ whenever $n \in N$ and $r \in R$.

Example 7.7.

(i) Both $\{0\}$ and M are submodules of a module M. A ***proper submodule*** of M is a submodule $N \subseteq M$ with $N \neq M$. In this case, we may write $N \subsetneq M$.

(ii) If a commutative ring R is viewed as a module over itself, then a submodule of R is an ideal; I is a proper submodule when it is a proper ideal.

(iii) A submodule of a $\mathbb{Z}$-module (i.e., of an abelian group) is a subgroup, and a submodule of a vector space is a subspace.

(iv) A submodule W of V^T, where $T : V \to V$ is a linear transformation, is a subspace W of V with $T(W) \subseteq W$ (it is clear that a submodule has this property; the converse is left as an exercise for the reader). Such a subspace is called an ***invariant subspace***.

(v) If M is an R-module and $r \in R$, then

$$rM = \{rm : m \in M\}$$

is a submodule of M.

Here is a related construction. If J is an ideal in R and M is an R-module, then

$$JM = \Big\{\sum_i j_i m_i : j_i \in J \text{ and } m_i \in M\Big\}$$

is a submodule of M.

(vi) If S and T are submodules of a module M, then

$$S + T = \{s + t : s \in S \text{ and } t \in T\}$$

is a submodule of M which contains S and T.

(vii) If $\{S_i : i \in I\}$ is a family of submodules of a module M, then $\bigcap_{i \in I} S_i$ is a submodule of M.

(viii) If M is an R-module and $m \in M$, then the ***cyclic submodule generated by*** m, denoted by $\langle m \rangle$, is

$$\langle m \rangle = \{rm : r \in R\}.$$

More generally, if X is a subset of an R-module M, then

$$\langle X \rangle = \Big\{\sum_{\text{finite}} r_i x_i : r_i \in R \text{ and } x_i \in X\Big\},$$

the set of all ***R-linear combinations*** of elements in X. We call $\langle X \rangle$ the ***submodule generated by*** X. See Exercise 7.2 on page 440. ◀

Definition. A module M is ***finitely generated*** if M is generated by a finite set; that is, if there is a finite subset $X = \{x_1, \ldots, x_n\}$ with $M = \langle X \rangle$.

For example, a vector space is finitely generated if and only if it is finite-dimensional.

We continue extending definitions from abelian groups and vector spaces to modules.

Definition. If $f : M \to N$ is an R-map between R-modules, then

$$\textbf{\textit{kernel}}\, \boldsymbol{f} = \ker f = \{m \in M : f(m) = 0\}$$

and

$$\textbf{\textit{image}}\, \boldsymbol{f} = \operatorname{im} f = \{n \in N : \text{there exists } m \in M \text{ with } n = f(m)\}.$$

It is routine to check that $\ker f$ is a submodule of M and that $\operatorname{im} f$ is a submodule of N. Suppose that $M = \langle X \rangle$; that is, M is generated by a subset X. Suppose further that N is a module and that $f, g : M \to N$ are R-homomorphisms. If f and g agree on X [that is, if $f(x) = g(x)$ for all $x \in X$], then $f = g$. The reason is that $f - g : M \to N$, defined by $f - g : m \mapsto f(m) - g(m)$, is an R-homomorphism with $X \subseteq \ker(f - g)$. Therefore, $M = \langle X \rangle \subseteq \ker(f - g)$, and so $f - g$ is identically zero; that is, $f = g$.

Definition. If N is a submodule of an R-module M, then the ***quotient module*** is the quotient group M/N (remember that M is an abelian group and N is a subgroup) equipped with the scalar multiplication

$$r(m + N) = rm + N.$$

The ***natural map***. $\pi : M \to M/N$, given by $m \mapsto m + N$, is easily seen to be an R-map.

Scalar multiplication in the definition of quotient module is well-defined: If $m + N = m' + N$, then $m - m' \in N$, hence $r(m - m') \in N$ (because N is a submodule), and so $rm - rm' \in N$ and $rm + N = rm' + N$.

Theorem 7.8 (First Isomorphism Theorem). *If $f : M \to N$ is an R-map of modules, then there is an R-isomorphism*

$$\varphi : M/\ker f \to \operatorname{im} f$$

given by

$$\varphi : m + \ker f \mapsto f(m).$$

Proof. If we view M and N only as abelian groups, then the first isomorphism theorem for groups says that $\varphi : M/\ker f \to \operatorname{im} f$ is an isomorphism of abelian groups. But φ is an R-map: $\varphi(r(m + N)) = \varphi(rm + N) = f(rm)$; since f is an R-map, however, $f(rm) = rf(m) = r\varphi(m + N)$, as desired. •

The second and third isomorphism theorems are corollaries of the first one.

Theorem 7.9 (Second Isomorphism Theorem). *If S and T are submodules of a module M, then there is an R-isomorphism*

$$S/(S \cap T) \to (S + T)/T.$$

Proof. Let $\pi : M \to M/T$ be the natural map, so that $\ker \pi = T$; define $h = \pi|S$, so that $h : S \to M/T$. Now

$$\ker h = S \cap T$$

and

$$\operatorname{im} h = (S + T)/T$$

[for $(S + T)/T$ consists of all those cosets in M/T having a representative in S]. The first isomorphism theorem now applies. •

Theorem 7.10 (Third Isomorphism Theorem). *If $T \subseteq S \subseteq M$ is a tower of submodules, then there is an R-isomorphism*

$$(M/T)/(S/T) \to M/S.$$

Proof. Define the map $g\colon M/T \to M/S$ to be ***coset enlargement***; that is,

$$g\colon m + T \mapsto m + S.$$

Now g is well-defined: If $m + T = m' + T$, then $m - m' \in T \subseteq S$ and $m + S = m' + S$. Moreover,

$$\ker g = S/T$$

and

$$\operatorname{im} g = M/S.$$

Again, the first isomorphism theorem completes the proof. •

If $f\colon M \to N$ is a map of modules and if $S \subseteq N$, then the reader may check that

$$f^{-1}(S) = \{m \in M\colon\ f(m) \in S\}$$

is a submodule of M containing $\ker f$.

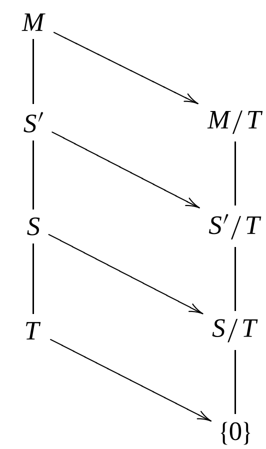
Theorem 7.11 (Correspondence Theorem). *If T is a submodule of a module M, then there is a bijection*

$$\varphi\colon \{\textit{intermediate submodules } T \subseteq S \subseteq M\} \to \{\textit{submodules of } M/T\}$$

given by

$$S \mapsto S/T.$$

Moreover, $S \subseteq S'$ in M if and only if $S/T \subseteq S'/T$ in M/T.



Proof. Since every module is an additive abelian group, every submodule is a subgroup, and so the correspondence theorem for groups, Theorem 2.76, shows that φ is an injection that preserves inclusions: $S \subseteq S'$ in M if and only if $S/T \subseteq S'/T$ in M/T. The remainder of this proof is a straightforward adaptation of the proof of Proposition 6.1; we need check only that additive homomorphisms are now R-maps. •

Proposition 7.12. *An R-module M is cyclic if and only if $M \cong R/I$ for some ideal I.*

Proof. If M is cyclic, then $M = \langle m \rangle$ for some $m \in M$. Define $f : R \to M$ by $f(r) = rm$. Now f is surjective, since M is cyclic, and its kernel is some ideal I. The first isomorphism theorem gives $R/I \cong M$.

Conversely, R/I is cyclic with generator $1 + I$, and any module isomorphic to a cyclic module is itself cyclic. •

Definition. A module M is ***simple*** (or ***irreducible***) if $M \neq \{0\}$ and M has no proper nonzero submodules; that is, the only submodules of M are $\{0\}$ and M.

Example 7.13.
By Proposition 2.107, an abelian group G is simple if and only if $G \cong \mathbb{I}_p$ for some prime p. ◄

Corollary 7.14. *An R-module M is simple if and only if $M \cong R/I$, where I is a maximal ideal.*

Proof. This follows from the correspondence theorem. •

Thus, the existence of maximal ideals guarantees the existence of simple modules.

The notion of direct sum, already discussed for vector spaces and for abelian groups, extends to modules. Recall that an abelian group G is an *internal* direct sum of subgroups S and T if $S + T = G$ and $S \cap T = \{0\}$, while an *external* direct sum is the abelian group whose underlying set is the cartesian product $S \times T$ and whose binary operation is pointwise addition; both versions give isomorphic abelian groups. The internal-external viewpoints persist for modules.

Definition. If S and T are R-modules, where R is a commutative[1] ring, then their ***direct sum***, denoted[2] by $S \sqcup T$, is the cartesian product $S \times T$ with coordinatewise operations:

$$(s, t) + (s', t') = (s + s', t + t');$$
$$r(s, t) = (rs, rt),$$

where $s, s' \in S$, $t, t' \in T$, and $r \in R$.

There are injective R-maps $\lambda_S \colon S \to S \sqcup T$ and $\lambda_T \colon T \to S \sqcup T$ given, respectively, by $\lambda_S \colon s \mapsto (s, 0)$ and $\lambda_T \colon t \mapsto (0, t)$.

Proposition 7.15. *The following statements are equivalent for R-modules M, S, and T.*

(i) $S \sqcup T \cong M$.

(ii) *There exist injective R-maps $i \colon S \to M$ and $j \colon T \to M$ such that*

$$M = \operatorname{im} i + \operatorname{im} j \qquad \text{and} \qquad \operatorname{im} i \cap \operatorname{im} j = \{0\}.$$

(iii) *There exist R-maps $i \colon S \to M$ and $j \colon T \to M$ such that, for every $m \in M$, there are unique $s \in S$ and $t \in T$ with*

$$m = is + jt.$$

(iv) *There are R-maps $i \colon S \to M$, $j \colon T \to M$, $p \colon M \to S$, and $q \colon M \to T$ such that*

$$pi = 1_S, \quad qj = 1_T, \quad pj = 0, \quad qi = 0, \; \textit{and} \quad ip + jq = 1_M.$$

Remark. The maps i and j are called ***injections***, and the maps p and q are called ***projections***. The equations $pi = 1_S$ and $qj = 1_T$ show that the maps i and j must be injective (so that $\operatorname{im} i \cong S$ and $\operatorname{im} j \cong T$) and the maps p and q must be surjective. ◂

[1] Modules over noncommutative rings are defined in the next chapter.

[2] Other common notations are $S \oplus T$ and $S \times T$.

Proof. (i) $\Rightarrow$ (ii) Let $\varphi\colon S \sqcup T \to M$ be an isomorphism, and define $i = \varphi\lambda_S$ [where $\lambda_S\colon s \mapsto (s, 0)$] and $j = \varphi\lambda_T$ [where $\lambda_T\colon t \mapsto (0, t)$]. Both i and j are injections, being the composites of injections. If $m \in M$, there is a unique ordered pair $(s, t) \in S \sqcup T$ with $m = \varphi((s, t))$. Hence,

$$m = \varphi((s, t)) = \varphi((s, 0) + (0, t)) = \varphi\lambda_S(s) + \varphi\lambda_T(t) = is + jt \in \operatorname{im} i + \operatorname{im} j.$$

If $x \in \operatorname{im} i \cap \operatorname{im} j$, then $is = jt$ for $s \in S$ and $t \in T$; that is, $\varphi\lambda_S(s) = \varphi\lambda_T(t)$. Since φ is an isomorphism, we have $(s, 0) = \lambda_S(s) = \lambda_T(t) = (0, t)$ in $S \sqcup T$. Therefore, $s = 0 = t$, $x = 0$, and $\operatorname{im} i \cap \operatorname{im} j = \{0\}$.

(ii) $\Rightarrow$ (iii) Given $m \in M$, an expression of the form $m = is + jt$ exists, by part (ii), and so we need prove only uniqueness. If also $m = is' + jt'$, then $i(s - s') = j(t' - t) \in \operatorname{im} i \cap \operatorname{im} j = \{0\}$. Therefore, $i(s - s') = 0$ and $j(t - t') = 0$. Since i and j are injections, we have $s = s'$ and $t = t'$.

(iii) $\Rightarrow$ (iv) If $m \in M$, then there are unique $s \in S$ and $t \in T$ with $m = is + jt$. The functions p and q, defined by

$$p(m) = s \qquad \text{and} \qquad q(m) = t,$$

are thus well-defined. It is routine to check that p and q are R-maps and that the first four equations in the statement hold (they follow from the definitions of p and q). For the last equation, if $m \in M$, then $m = is + jt$, and $ip(m) + jq(m) = is + jt = m$.

(iv) $\Rightarrow$ (i) Define $\varphi\colon S \sqcup T \to M$ by $\varphi\colon (s, t) \mapsto is + jt$. It is easy to see that φ is an R-map; φ is surjective because $1_M = ip + jq$. To see that φ is injective, suppose that $\varphi((s, t)) = 0$, so that $is = -jt$. Now $s = pis = -pjt = 0$ and $-t = -qjt = qis = 0$, as desired. •

Internal direct sum is probably the most important instance of a module isomorphic to a direct sum.

Definition. If S and T are submodules of a module M, then M is their ***internal direct sum*** if $M \cong S \sqcup T$ with $i\colon S \to M$ and $j\colon T \to M$ the inclusions. We denote an internal direct sum by

$$M = S \oplus T.$$

In this chapter only, we will use the notation $S \sqcup T$ to denote the external direct sum (underlying set the cartesian product of all ordered pairs) and the notation $M = S \oplus T$ to denote the internal direct sum (S and T submodules of M as just defined). Later, we shall write as the mathematical world writes: The same notation $S \oplus T$ is used for either version of direct sum.

Here is a restatement of Proposition 7.15 for internal direct sums.

Corollary 7.16. *The following conditions are equivalent for an R-module M with submodules S and T.*

(i) $M = S \oplus T$.

(ii) *$S + T = M$ and $S \cap T = \{0\}$.*

(iii) *Each $m \in M$ has a unique expression of the form $m = s + t$ for $s \in S$ and $t \in T$.*

Proof. This follows at once from Proposition 7.15 by taking i and j to be inclusions. •

Definition. A submodule S of a module M is a ***direct summand*** of M if there exists a submodule T of M with $M = S \oplus T$.

The next corollary will connect direct summands with a special type of homomorphism.

Definition. If S is a submodule of an R-module M, then S is a ***retract*** of M if there exists an R-homomorphism $\rho\colon M \to S$, called a ***retraction***, with $\rho(s) = s$ for all $s \in S$.

Retractions in nonabelian groups arose in Exercise 5.72 on page 318.

Corollary 7.17. *A submodule S of a module M is a direct summand if and only if there exists a retraction $\rho\colon M \to S$.*

Proof. In this case, we let $i\colon S \to M$ be the inclusion. We show that $M = S \oplus T$, where $T = \ker \rho$. If $m \in M$, then $m = (m - \rho m) + \rho m$. Plainly, $\rho m \in \operatorname{im} \rho = S$. On the other hand, $\rho(m - \rho m) = \rho m - \rho\rho m = 0$, because $\rho m \in S$ and so $\rho\rho m = \rho m$. Therefore, $M = S + T$.

If $m \in S$, then $\rho m = m$; if $m \in T = \ker \rho$, then $\rho m = 0$. Hence, if $m \in S \cap T$, then $m = 0$. Therefore, $S \cap T = \{0\}$, and $M = S \oplus T$.

For the converse, if $M = S \oplus T$, then each $m \in M$ has a unique expression of the form $m = s + t$, where $s \in S$ and $t \in T$, and it is easy to check that $\rho\colon M \to S$, defined by $\rho\colon s + t \mapsto s$, is a retraction $M \to S$. •

Corollary 7.18. *If $M = S \oplus T$ and $S \subseteq A \subseteq M$, then $A = S \oplus (A \cap T)$.*

Proof. Let $\rho\colon M \to S$ be the retraction $s + t \mapsto s$. Since $S \subseteq A$, the restriction $\rho|A\colon A \to S$ is a retraction with $\ker \rho|A = A \cap T$. •

The direct sum construction can be extended to finitely many submodules. There is an external and internal version.

Definition. Let $S_1, \ldots, S_n$ be R-modules. Define the ***external direct sum***

$$S_1 \sqcup \cdots \sqcup S_n$$

to be the R-module whose underlying set is the cartesian product $S_1 \times \cdots \times S_n$ and whose operations are

$$(s_1, \ldots, s_n) + (s_1', \ldots, s_n') = (s_1 + s_1', \ldots, s_n + s_n')$$
$$r(s_1, \ldots, s_n) = (rs_1, \ldots, rs_n).$$

Let M be a module, and let $S_1, \ldots, S_n$ be submodules of M. Define M to be the ***internal direct sum***

$$M = S_1 \oplus \cdots \oplus S_n$$

if each $m \in M$ has a unique expression of the form $m = s_1 + \cdots + s_n$, where $s_i \in S_i$ for all $i = 1, \ldots, n$.

We let the reader prove that both internal and external versions, when the former is defined, are isomorphic.

For example, if V is an n-dimensional vector space over a field k, and if $v_1, \ldots, v_n$ is a basis, then

$$V = \langle v_1 \rangle \oplus \cdots \oplus \langle v_n \rangle.$$

If $S_1, \ldots, S_n$ are submodules of a module M, when is $\langle S_1, \ldots, S_n \rangle$, the submodule generated by the S_i, equal to their direct sum? A common mistake is to say that it is enough to assume that $S_i \cap S_j = \{0\}$ for all $i \neq j$, but Example 5.3 on page 251 shows that this is not enough.

Proposition 7.19. *Let $M = S_1 + \cdots + S_n$, where the S_i are submodules; that is, each $m \in M$ has a (not necessarily unique) expression of the form*

$$m = s_1 + \cdots + s_n,$$

where $s_i \in S_i$ for all i. Then $M = S_1 \oplus \cdots \oplus S_n$ if and only if, for each i,

$$S_i \cap \langle S_1 + \cdots + \widehat{S_i} + \cdots + S_n \rangle = \{0\},$$

where $\widehat{S_i}$ means that the term S_i is omitted from the sum.

Proof. A straightforward adaptation of Proposition 5.4. See Exercise 7.79 on page 519 for the generalization of this proposition for infinitely many submodules. •

Here is the last definition in this dictionary of modules.

Definition. A sequence of R-maps and R-modules

$$\cdots \to M_{n+1} \xrightarrow{f_{n+1}} M_n \xrightarrow{f_n} M_{n-1} \to \cdots$$

is called an ***exact sequence***[3] if $\operatorname{im} f_{n+1} = \ker f_n$ for all n.

Observe that there is no need to label an arrow $0 \xrightarrow{f} A$ or $B \xrightarrow{g} 0$ for, in either case, there is a unique map, namely, $f : 0 \mapsto 0$ or the constant homomorphism $g(b) = 0$ for all $b \in B$.[4]

Here are some simple consequences of a sequence of homomorphisms being exact.

[3]This terminology comes from advanced calculus, where a differential form ω is called ***closed*** if $d\omega = 0$ and it is called ***exact*** if $\omega = dh$ for some function h (see Proposition 9.146 on page 753). The term was coined by the algebraic topologist W. Hurewicz. It is interesting to look at the book by Hurewicz–Wallman, *Dimension Theory*, which was written just before this coinage. We can see there many results that would have been much simpler to state had the word *exact* been available.

[4]In diagrams, we usually write 0 instead of $\{0\}$.

Proposition 7.20.

(i) *A sequence* $0 \to A \xrightarrow{f} B$ *is exact if and only if* f *is injective.*

(ii) *A sequence* $B \xrightarrow{g} C \to 0$ *is exact if and only if* g *is surjective.*

(iii) *A sequence* $0 \to A \xrightarrow{h} B \to 0$ *is exact if and only if* h *is an isomorphism.*

Proof. (i) The image of $0 \to A$ is $\{0\}$, so that exactness gives $\ker f = \{0\}$, and so f is injective. Conversely, given $f\colon A \to B$, there is an exact sequence $\ker f \to A \xrightarrow{f} B$. If f is injective, then $\ker f = \{0\}$.

(ii) The kernel of $C \to 0$ is C, so that exactness gives $\operatorname{im} g = C$, and so g is surjective. Conversely, given $g\colon B \to C$, there is an exact sequence $B \xrightarrow{g} C \to C/\operatorname{im} g$ (see Exercise 7.13). If g is surjective, then $C = \operatorname{im} g$ and $C/\operatorname{im} g = \{0\}$.

(iii) Part (i) shows that h is injective if and only if $0 \to A \xrightarrow{h} B$ is exact; part (ii) shows that h is surjective if and only if $A \xrightarrow{h} B \to 0$ is exact. Therefore, h is an isomorphism if and only if the sequence $0 \to A \xrightarrow{h} B \to 0$ is exact. •

We can restate the isomorphism theorems in the language of exact sequences.

Definition. A ***short exact sequence*** is an exact sequence of the form

$$0 \to A \xrightarrow{f} B \xrightarrow{g} C \to 0.$$

We also call this short exact sequence an ***extension*** of A by C.

Some authors call this an extension of C by A; some authors say that the middle module B is an extension.

Proposition 7.21.

(i) *If* $0 \to A \xrightarrow{f} B \xrightarrow{g} C \to 0$ *is a short exact sequence, then*

$$A \cong \operatorname{im} f \qquad \textit{and} \qquad B/\operatorname{im} f \cong C.$$

(ii) *If* $T \subseteq S \subseteq M$ *is a tower of submodules, then there is an exact sequence*

$$0 \to S/T \xrightarrow{f} M/T \xrightarrow{g} M/S \to 0.$$

Proof. (i) Since f is injective, it is an isomorphism $A \to \operatorname{im} f$. The first isomorphism theorem gives $B/\ker g \cong \operatorname{im} g$. By exactness, however, $\ker g = \operatorname{im} f$ and $\operatorname{im} g = C$; therefore, $B/\operatorname{im} f \cong C$.

(ii) This is just a restatement of the third isomorphism theorem. Define $f\colon S/T \to M/T$ to be the inclusion, and define $g\colon M/T \to M/S$ be "coset enlargement:" $g\colon m + T \mapsto m + S$. As in the proof of Theorem 7.10, g is surjective, and $\ker g = S/T = \operatorname{im} f$. •

In the special case when A is a submodule of B and $f : A \to B$ is the inclusion, then exactness of $0 \to A \xrightarrow{f} B \xrightarrow{g} C \to 0$ gives $B/A \cong C$.

Definition. A short exact sequence

$$0 \to A \xrightarrow{i} B \xrightarrow{p} C \to 0$$

is ***split*** if there exists a map $j : C \to B$ with $pj = 1_C$.

Proposition 7.22. *If an exact sequence*

$$0 \to A \xrightarrow{i} B \xrightarrow{p} C \to 0$$

is split, then $B \cong A \sqcup C$.

Remark. Exercise 7.17 on page 441 characterizes split short exact sequences. ◂

Proof. We show that $B = \operatorname{im} i \oplus \operatorname{im} j$, where $j : C \to B$ satisfies $pj = 1_C$. If $b \in B$, then $pb \in C$ and $b - jpb \in \ker p$, for $p(b - jpb) = pb - pj(pb) = 0$ because $pj = 1_C$. By exactness, there is $a \in A$ with $ia = b - jpb$. It follows that $B = \operatorname{im} i + \operatorname{im} j$. It remains to prove $\operatorname{im} i \cap \operatorname{im} j = \{0\}$. If $ia = x = jc$, then $px = pia = 0$, because $pi = 0$, whereas $px = pjc = c$, because $pj = 1_C$. Therefore, $x = jc = 0$, and so $B \cong A \sqcup C$. •

The converse of the last proposition is not true. Let $A = \langle a \rangle$, $B = \langle b \rangle$, and $C = \langle c \rangle$ be cyclic groups of orders 2, 4, and 2, respectively. If $i : A \to B$ is defined by $i(a) = 2b$ and $p : B \to C$ is defined by $p(b) = c$, then $0 \to A \xrightarrow{i} B \xrightarrow{p} C \to 0$ is an exact sequence which is not split: $\operatorname{im} i = \langle 2b \rangle$ is not even a pure subgroup of B. By Exercise 7.12 on page 440, for any abelian group M, there is an exact sequence

$$0 \to A \xrightarrow{i'} B \sqcup M \xrightarrow{p'} C \sqcup M \to 0,$$

where $i'(a) = (2b, 0)$ and $p'(b, m) = (c, m)$, and this sequence does not split either. If we choose $M = \mathbb{I}_4[x] \sqcup \mathbb{I}_2[x]$ (the direct summands are the polynomial rings over $\mathbb{I}_4$ and $\mathbb{I}_2$, respectively), then $A \sqcup (C \sqcup M) \cong B \sqcup M$. (For readers who are familiar with infinite direct sums, which we introduce later in this chapter, M is the direct sum of infinitely many copies of $\mathbb{I}_4 \sqcup \mathbb{I}_2$.)

Here is a characterization of noetherian rings using these ideas.

Proposition 7.23.

(i) *A commutative ring R is noetherian if and only if every submodule of a finitely generated R-module M is itself finitely generated*

(ii) *If R is a* PID *and if M can be generated by n elements, then every submodule of M can be generated by n or fewer elements.*

Remark. Proposition 7.23(ii) is not true more generally. For example, if R is not a PID, there there is some ideal I that is not principal. Thus, R has one generator while its submodule I cannot be generated by one element. ◂

Proof. (i) Assume that every submodule of a finitely generated R-module is finitely generated. In particular, every submodule of R, which is a cyclic R-module and hence finitely generated, is finitely generated. But submodules of R are ideals, and so every ideal is finitely generated; that is, R is noetherian.

We prove the converse by induction on $n \geq 1$, where $M = \langle x_1, \ldots, x_n \rangle$. If $n = 1$, then M is cyclic, and so Proposition 7.12 gives $M \cong R/I$ for some ideal I. If $S \subseteq M$, then the correspondence theorem gives an ideal J with $I \subseteq J \subseteq R$ and $S \cong J/I$. But R is noetherian, so that J, and hence J/I, is finitely generated

If $n \geq 1$ and $M = \langle x_1, \ldots, x_n, x_{n+1} \rangle$, consider the exact sequence

$$0 \to M' \xrightarrow{i} M \xrightarrow{p} M'' \to 0,$$

where $M' = \langle x_1, \ldots, x_n \rangle$, $M'' = M/M'$, i is the inclusion, and p is the natural map. Note that M'' is cyclic, being generated by $x_{n+1} + M'$. If $S \subseteq M$ is a submodule, there is an exact sequence

$$0 \to S \cap M' \to S \to S/(S \cap M') \to 0.$$

Now $S \cap M' \subseteq M'$, and hence it is finitely generated, by the inductive hypothesis. Furthermore, $S/(S \cap M') \cong (S + M')/M' \subseteq M/M'$, so that $S/(S \cap M')$ is finitely generated, by the base step. Using Exercise 7.15 on page 441, we conclude that S is finitely generated

(ii) We prove the statement by induction on $n \geq 1$. If M is cyclic, then $M \cong R/I$; if $S \subseteq M$, then $S \cong J/I$ for some ideal J in R containing I. Since R is a PID, J is principal, and so J/I is cyclic.

For the inductive step, we refer to the exact sequence

$$0 \to S \cap M' \to S \to S/(S \cap M') \to 0$$

in part (i), where $M = \langle x_1, \ldots, x_n, x_{n+1} \rangle$ and $M' = \langle x_1, \ldots, x_n \rangle$. By the inductive hypothesis, $S \cap M'$ can be generated by n or fewer elements, while the base step shows that $S/(S \cap M')$ is cyclic. Exercise 7.15 on page 441 shows that S can be generated by $n + 1$ or fewer elements. •

The next proposition, whose proof uses Proposition 7.23(ii), shows that the sum and product of algebraic integers are themselves algebraic integers. If α and β are algebraic integers, it is not too difficult to give monic polynomials having $\alpha + \beta$ and $\alpha\beta$ as roots, but it takes a bit of work to find such polynomials having all coefficients in $\mathbb{Z}$ (see Pollard, *The Theory of Algebraic Numbers*, page 33).

Proposition 7.24. *Let* $\alpha \in \mathbb{C}$ *and define* $\mathbb{Z}[\alpha] = \{g(\alpha) : g(x) \in \mathbb{Z}[x]\}$.

(i) $\mathbb{Z}[\alpha]$ *is a subring of* $\mathbb{C}$.

(ii) *A complex number α is an algebraic integer if and only if $\mathbb{Z}[\alpha]$ is a finitely generated additive abelian group.*

(iii) *The set of all the algebraic integers is a subring of $\mathbb{C}$.*

Proof. (i) Since $1 = g(\alpha)$, where $g(x) = 1$ is a constant polynomial, we have $1 \in \mathbb{Z}[\alpha]$. If $f(\alpha), g(\alpha) \in \mathbb{Z}[\alpha]$, then so is $f(\alpha) + g(\alpha) = h(\alpha)$, where $h(x) = f(x) + g(x)$. Similarly, $f(\alpha)g(\alpha) \in \mathbb{Z}[\alpha]$, and so $\mathbb{Z}[\alpha]$ is a subring of $\mathbb{C}$.

(ii) If α is an algebraic integer, there is a monic polynomial $f(x) \in \mathbb{Z}[x]$ having α as a root. We claim that if $\deg(f) = n$, then $\mathbb{Z}[\alpha] = G$, where G is the set of all linear combinations $m_0 + m_1\alpha + \cdots + m_{n-1}\alpha^{n-1}$ with $m_i \in \mathbb{Z}$. Clearly, $G \subseteq \mathbb{Z}[\alpha]$. For the reverse inclusion, each element $u \in \mathbb{Z}[\alpha]$ has the form $u = g(\alpha)$, where $g(x) \in \mathbb{Z}[x]$. Since $f(x)$ is monic, the division algorithm (Corollary 3.22) gives $q(x), r(x) \in \mathbb{Z}[x]$ with $g(x) = q(x) f(x) + r(x)$, where either $r(x) = 0$ or $\deg(r) < \deg(f) = n$. Therefore,

$$u = g(\alpha) = q(\alpha) f(\alpha) + r(\alpha) = r(\alpha) \in G.$$

Thus, the additive group of $\mathbb{Z}[\alpha]$ is finitely generated.

Conversely, if the additive group of the commutative ring $\mathbb{Z}[\alpha]$ is finitely generated, that is, $\mathbb{Z}[\alpha] = \langle g_1, \ldots, g_m \rangle$ as an abelian group, then each g_j is a $\mathbb{Z}$-linear combination of powers of α. Let m be the largest power of α occurring in any of these g's. Since $\mathbb{Z}[\alpha]$ is a commutative ring, $\alpha^{m+1} \in \mathbb{Z}[\alpha]$; hence, α^{m+1} can be expressed as a $\mathbb{Z}$-linear combination of smaller powers of α; say, $\alpha^{m+1} = \sum_{i=0}^{m} b_i\alpha^i$, where $b_i \in \mathbb{Z}$. Therefore, α is a root of $f(x) = x^{m+1} - \sum_{i=0}^{m} b_i x^i$, which is a monic polynomial in $\mathbb{Z}[x]$, and so α is an algebraic integer.

(iii) Suppose that α and β are algebraic integers; let α be a root of a monic $f(x) \in \mathbb{Z}[x]$ of degree n, and let β be a root of a monic $g(x) \in \mathbb{Z}[x]$ of degree m. Now $\mathbb{Z}[\alpha\beta]$ is an additive subgroup of $G = \langle \alpha^i \beta^j : 0 \le i < n, 0 \le j < m \rangle$. Since G is finitely generated, so is its subgroup $\mathbb{Z}[\alpha\beta]$, by Proposition 7.23(ii), and so $\alpha\beta$ is an algebraic integer. Similarly, $\mathbb{Z}[\alpha + \beta]$ is an additive subgroup of $\langle \alpha^i \beta^j : i + j \le n + m - 1 \rangle$, and so $\alpha + \beta$ is also an algebraic integer. •

This last theorem gives a technique for proving that an integer a is a divisor of an integer b. If we can prove that b/a is an algebraic integer, then it must be an integer, for it is obviously rational. This will actually be used in Chapter 8 to prove that the degrees of the irreducible characters of a finite group G are divisors of $|G|$.

EXERCISES

7.1 Let R be a commutative ring. Call an (additive) abelian group M an ***almost R-module*** if there is a function $R \times M \to M$ satisfying all the axioms of an R-module except axiom (iv): We do not assume that $1m = m$ for all $m \in M$.

Prove that

$$M = M_1 \oplus M_0,$$

where

$$M_1 = \{m \in M\colon 1m = m\} \text{ and } M_0 = \{m \in M\colon rm = 0 \text{ for all } r \in R\}$$

are subgroups of M that are almost R-modules; in fact, M_1 is an R-module.

7.2 If X is a subset of a module M, prove that $\langle X\rangle$, the submodule of M generated by X, is equal to $\bigcap S$, where the intersection ranges over all those submodules $S \subseteq M$ containing X.

7.3 Prove that if $f\colon M \to N$ is an R-map and K is a submodule of M with $K \subseteq \ker f$, then f induces an R-map $\overline{f}\colon M/K \to N$ by $\overline{f}\colon m + K \mapsto f(m)$.

7.4 Let R be a commutative ring and let J be an ideal in R. Recall that if M is an R-module, then $JM = \left\{\sum_i j_i m_i\colon j_i \in J \text{ and } m_i \in M\right\}$ is a submodule of M. Prove that M/JM is an R/J-module if we define scalar multiplication:

$$(r + J)(m + JM) = rm + JM.$$

Conclude that if $JM = \{0\}$, then M itself is an R/J-module; in particular, if J is a maximal ideal in R and $JM = \{0\}$, then M is a vector space over R/J.

7.5 For every R-module M, prove that there is an R-isomorphism

$$\varphi_M\colon \operatorname{Hom}_R(R, M) \to M,$$

given by $\varphi_M\colon f \mapsto f(1)$.

7.6 Let $F = \sum_{i=1}^n \langle b_i\rangle$ be a direct sum of R-modules, where $f_i\colon R \to \langle b_i\rangle$, given by $r \mapsto rb_i$, is an isomorphism. Prove that if M is a maximal ideal in R, then the cosets $\{b_i + MF\colon i = 1, \ldots, n\}$ form a basis of the vector space F/MF over the field R/M. (See Exercise 7.4.)

7.7 Let R and S be commutative rings, and let $\varphi\colon R \to S$ be a ring homomorphism. If M is an S-module, prove that M is also an R-module if we define

$$rm = \varphi(r)m,$$

for all $r \in R$ and $m \in M$.

7.8 Let $M = S_1 \sqcup \cdots \sqcup S_n$ be a direct sum of R-modules. If $T_i \subseteq S_i$ for all i, prove that

$$(S_1 \sqcup \cdots \sqcup S_n)/(T_1 \sqcup \cdots \sqcup T_n) \cong (S_1/T_1) \sqcup \cdots \sqcup (S_n/T_n).$$

7.9 Let R be a commutative ring and let M be a nonzero R-module. If $m \in M$, define $\operatorname{ord}(m) = \{r \in R : rm = 0\}$, and define $\mathcal{F} = \{\operatorname{ord}(m) : m \in M \text{ and } m \neq 0\}$. Prove that every maximal element in $\mathcal{F}$ is a prime ideal.

7.10 Let $A \xrightarrow{f} B \xrightarrow{g} C$ be a sequence of module maps. Prove that $gf = 0$ if and only if $\operatorname{im} f \subseteq \ker g$. Give an example of such a sequence that is not exact.

7.11 If $0 \to M \to 0$ is an exact sequence, prove that $M = \{0\}$.

7.12 Let $0 \to A \to B \to C \to 0$ be a short exact sequence of modules. If M is any module, prove that there are exact sequences

$$0 \to A \oplus M \to B \oplus M \to C \to 0$$

and

$$0 \to A \to B \oplus M \to C \oplus M \to 0.$$

Definition. If $f : M \to N$ is a map, define its ***cokernel***, denoted by coker f, as

$$\operatorname{coker} f = N/\operatorname{im} f.$$

7.13 (i) Prove that a map $f : M \to N$ is surjective if and only if coker $f = \{0\}$.

(ii) If $f : M \to N$ is a map, prove that there is an exact sequence

$$0 \to \ker f \to M \xrightarrow{f} N \to \operatorname{coker} f \to 0.$$

7.14 If $A \xrightarrow{f} B \to C \xrightarrow{h} D$ is an exact sequence, prove that f is surjective if and only if h is injective.

7.15 Let $0 \to A \xrightarrow{i} B \xrightarrow{p} C \to 0$ be a short exact sequence.

(i) Assume that $A = \langle X \rangle$ and $C = \langle Y \rangle$. For each $y \in Y$, choose $y' \in B$ with $p(y') = y$. Prove that

$$B = \big\langle i(X) \cup \{y' : y \in Y\} \big\rangle.$$

(ii) Prove that if both A and C are finitely generated, then B is finitely generated More precisely, prove that if A can be generated by m elements and if C can be generated by n elements, then B can be generated by $m + n$ elements.

7.16 Prove that every short exact sequence of vector spaces is split.

7.17 Prove that a short exact sequence

$$0 \to A \xrightarrow{i} B \xrightarrow{p} C \to 0$$

splits if and only if there exists $q : B \to A$ with $qi = 1_A$.

7.18 (i) Prove that a map $\varphi : B \to C$ is injective if and only if φ can be canceled from the left; that is, for all modules A and all maps $f, g : A \to B$, we have $\varphi f = \varphi g$ implies $f = g$.

$$A \underset{g}{\overset{f}{\rightrightarrows}} B \xrightarrow{\varphi} C$$

(ii) Prove that a R-map $\varphi : B \to C$ is surjective if and only if φ can be canceled from the right; that is, for all R-modules D and all R-maps $h, k : C \to D$, we have $h\varphi = k\varphi$ implies $h = k$.

$$B \xrightarrow{\varphi} C \underset{k}{\overset{h}{\rightrightarrows}} D$$

7.19 (***Eilenberg–Moore***) Let G be a (possibly nonabelian) group.

(i) If H is a proper subgroup of a group G, prove that there exists a group L and distinct homomorphisms $f, g : G \to L$ with $f|H = g|H$.

Hint. Define $L = S_X$, where X denotes the family of all the left cosets of H in G together with an additional element, denoted ∞. If $a \in G$, define $f(a) = f_a \in S_X$ by $f_a(\infty) = \infty$ and $f_a(bH) = abH$. Define $g : G \to S_X$ by $g = \gamma \circ f$, where $\gamma \in S_X$ is conjugation by the transposition (H, ∞).

(ii) If A and G are groups, prove that a homomorphism $\varphi : A \to G$ is surjective if and only if φ can be canceled from the right; that is, for all groups L and all maps $f, g : G \to L$, we have $f\varphi = g\varphi$ implies $f = g$.

$$B \xrightarrow{\varphi} G \underset{g}{\overset{f}{\rightrightarrows}} L$$

7.2 CATEGORIES

Imagine a set theory whose primitive terms, instead of *set* and *element*, are *set* and *function*. How could we define bijection, cartesian product, union, and intersection? Category theory will force us to think in this way. Now categories are the context for discussing general properties of systems such as groups, rings, vector spaces, modules, sets, and topological spaces, in tandem with their respective transformations: homomorphisms, functions, and continuous maps. There are two basic reasons for studying categories: The first is that they are needed to define functors and natural transformations (which we will do in the next sections); the other is that categories will force us to regard a module, for example, not in isolation, but in a context serving to relate it to all other modules (for example, we will define certain modules as solutions to *universal mapping problems*).

There are well-known set-theoretic "paradoxes" that show that contradictions arise if we are not careful about how the undefined terms *set* and *element* are used. For example, Russell's paradox shows how we can run into trouble by regarding every collection as a set. Define a ***Russell set*** to be a set S that is not a member of itself; that is, $S \notin S$. If R is the family of all Russell sets, is R a Russell set? On the one hand, if $R \in R$, then R is not a Russell set; as only Russell sets are members of R, we must have $R \notin R$, and this is a contradiction. On the other hand, if we assume that $R \notin R$, then R is a Russell set, and so it belongs to R (which contains every Russell set); again, we have a contradiction. We conclude that we must impose some conditions on what collections are allowed to be sets (and also some conditions on the membership relation $\in$). One way to avoid such problems is to axiomatize set theory by considering *class* as a primitive term instead of *set*. The axioms give the existence of finite classes and of $\mathbb{N}$; they also provide rules for constructing special classes from given ones, and any class constructed according to these rules is called a ***set***. Cardinality can be defined, and there is a theorem that a class is a set if and only if it is "small"; that is, it has a cardinal number. A ***proper class*** is defined to be a class that is not a set. For example, $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$, and $\mathbb{C}$ are sets, while the collection of all sets is a proper class. Paradoxes are avoided by decreeing that some rules apply only to sets but not to proper classes.

Definition. A ***category*** $\mathcal{C}$ consists of three ingredients: a class $\text{obj}(\mathcal{C})$ of ***objects***, a *set* of ***morphisms*** $\text{Hom}(A, B)$ for every ordered pair (A, B) of objects, and ***composition*** $\text{Hom}(A, B) \times \text{Hom}(B, C) \to \text{Hom}(A, C)$, denoted by

$$(f, g) \mapsto gf,$$

for every ordered triple A, B, C of objects. [We often write $f : A \to B$ or $A \xrightarrow{f} B$ to denote $f \in \text{Hom}(A, B)$.] These ingredients are subject to the following axioms:

(i) the Hom sets are pairwise disjoint;[5] that is, each morphism has a unique domain and a unique target;

[5]One can force pairwise disjointness by labeling morphisms $f \in \text{Hom}(A, B)$ by ${}_A f_B$.

(ii) for each object A, there is an ***identity morphism*** $1_A \in \text{Hom}(A, A)$ such that

$$f 1_A = f \text{ and } 1_B f = f \text{ for all } f : A \to B;$$

(iii) composition is associative: Given morphisms

$$A \xrightarrow{f} B \xrightarrow{g} C \xrightarrow{h} D,$$

then

$$h(gf) = (hg)f.$$

The important notion, in this circle of ideas, is not category but functor, which will be introduced in the next section. Categories are necessary because they are an essential ingredient in the definition of functor. A similar situation occurs in linear algebra: Linear transformation is the important notion, but we must first consider vector spaces in order to define it.

The following examples will explain certain fine points of the definition of category.

Example 7.25.

(i) $\mathcal{C} = \mathbf{Sets}$.

The objects in this category are sets (not proper classes), morphisms are functions, and composition is the usual composition of functions.

A standard result of set theory is that if A and B are sets, then $\text{Hom}(A, B)$, the class of all functions from A to B, is a set. That Hom sets are pairwise disjoint is just the reflection of the definition of equality of functions given in Chapter 1: In order that two functions be equal, they must, first, have the same domains and the same targets (and, of course, they must have the same graphs).

(ii) $\mathcal{C} = \mathbf{Groups}$.

Here, objects are groups, morphisms are homomorphisms, and composition is the usual composition (homomorphisms are functions).

(iii) $\mathcal{C} = \mathbf{CommRings}$.

Here, objects are commutative rings, morphisms are ring homomorphisms, and composition is the usual composition.

(iv) $\mathcal{C} = {}_R\mathbf{Mod}$.[6]

The objects in this category are R-modules, where R is a commutative ring, morphisms are R-homomorphisms, and composition is the usual composition. We denote the sets $\text{Hom}(A, B)$ in ${}_R\mathbf{Mod}$ by

$$\text{Hom}_R(A, B).$$

If $R = \mathbb{Z}$, then we often write

$${}_{\mathbb{Z}}\mathbf{Mod} = \mathbf{Ab}$$

to remind ourselves that $\mathbb{Z}$-modules are just abelian groups.

---

[6]When we introduce noncommutative rings in the Chapter 8, then we will denote the category of left R-modules by ${}_R\mathbf{Mod}$ and the category of right R-modules by $\mathbf{Mod}_R$.

(v) $\mathcal{C} = \mathbf{PO}(X)$.

If X is a partially ordered set, regard it as a category whose objects are the elements of X, whose Hom sets are either empty or have only one element:

$$\mathrm{Hom}(x, y) = \begin{cases} \varnothing & \text{if } x \not\preceq y \\ \kappa_y^x & \text{if } x \preceq y \end{cases}$$

(the symbol κ_y^x denotes the unique element in the Hom set when $x \preceq y$) and whose composition is given by

$$\kappa_z^y \kappa_y^x = \kappa_z^x.$$

Note that $1_x = \kappa_x^x$, by reflexivity, while composition makes sense because $\preceq$ is transitive.[7]

We insisted, in the definition of category, that $\mathrm{Hom}(A, B)$ be a set, but we left open the possibility that it be empty. The category $\mathbf{PO}(X)$ is an example in which this possibility occurs. [Not every Hom set in a category $\mathcal{C}$ can be empty, for $\mathrm{Hom}(A, A) \neq \varnothing$ for every object $A \in \mathcal{C}$ because it contains the identity morphism 1_A.]

(vi) $\mathcal{C} = \mathcal{C}(G)$.

If G is a group, then the following description defines a category $\mathcal{C}(G)$: There is only one object, denoted by $*$, $\mathrm{Hom}(*, *) = G$, and composition

$$\mathrm{Hom}(*, *) \times \mathrm{Hom}(*, *) \to \mathrm{Hom}(*, *);$$

that is, $G \times G \to G$, is the given multiplication in G. We leave verification of the axioms to the reader.[8]

The category $\mathcal{C}(G)$ has an unusual property. Since $*$ is merely an object, not a set, there are no *functions* $* \to *$ defined on it; thus, morphisms here are not functions. Another curious property of this category is another consequence of there being only one object: there are no proper subobjects here.

(vii) There are many interesting nonalgebraic examples of categories. For example, $\mathcal{C} =$ **Top**, the category with objects all topological spaces, morphisms all continuous functions, and usual composition. ◀

Here is how to translate *isomorphism* into categorical language.

Definition. A morphism $f : A \to B$ in a category $\mathcal{C}$ is an ***equivalence*** (or an ***isomorphism***) if there exists a morphism $g : B \to A$ in $\mathcal{C}$ with

$$gf = 1_A \quad \text{and} \quad fg = 1_B.$$

The morphism g is called the ***inverse*** of f.

---

[7] A nonempty set X is called ***quasi-ordered*** if it has a relation $x \preceq y$ that is reflexive and transitive (if, in addition, this relation is antisymmetric, then X is partially ordered). $\mathbf{PO}(X)$ is a category for every quasi-ordered set.

[8] That every element in G have an inverse is not needed to prove that $\mathcal{C}(G)$ is a category, and $\mathcal{C}(G)$ is a category for every monoid G.

It is easy to see that an inverse of an equivalence is unique.

Identity morphisms in a category are always equivalences. If $\mathcal{C} = \mathbf{PO}(X)$, where X is a partially ordered set, then the only equivalences are identities; if $\mathcal{C} = \mathcal{C}(G)$, where G is a group (see Example 7.25(vi)), then every morphism is an equivalence. If $\mathcal{C} = \mathbf{Sets}$, then equivalences are bijections; if $\mathcal{C} = \mathbf{Groups}$, $\mathcal{C} = {}_R\mathbf{Mod}$, or $\mathcal{C} = \mathbf{CommRings}$, then equivalences are isomorphisms; if $\mathcal{C} = \mathbf{Top}$, then equivalences are homeomorphisms.

Let us give a name to a feature of the category ${}_R\mathbf{Mod}$ (which we saw in Proposition 7.5) that is not shared by more general categories: Homomorphisms can be added.

Definition. A category $\mathcal{C}$ is ***pre-additive*** if every $\operatorname{Hom}(A, B)$ is equipped with a binary operation making it an (additive) abelian group for which the distributive laws hold: for all $f, g \in \operatorname{Hom}(A, B)$,

(i) if $p \colon B \to B'$, then

$$p(f + g) = pf + pg \in \operatorname{Hom}(A, B');$$

(ii) if $q \colon A' \to A$, then

$$(f + g)q = fq + gq \in \operatorname{Hom}(A', B).$$

In Exercise 7.22 on page 458, it is shown that **Groups** does not have the structure of a pre-additive category.

A category is defined in terms of objects and morphisms; its objects need not be sets, and its morphisms need not be functions [$\mathcal{C}(G)$ in Example 7.25(vi) is such a category]. We now give ourselves the exercise of trying to describe various constructions in **Sets** or in ${}_R\mathbf{Mod}$ so that they make sense in arbitrary categories.

In Proposition 7.15(iii), we gave the following characterization of direct sum $M = A \oplus B$: there are homomorphisms $p \colon M \to A$, $q \colon M \to B$, $i \colon A \to M$, and $j \colon B \to M$ such that

$$pi = 1_A,\ qj = 1_B,\ pj = 0,\ qi = 0 \quad \text{and} \quad ip + jq = 1_M.$$

Even though this description of direct sum is phrased in terms of arrows, it is not general enough to make sense in every category; it makes use of a property of the category ${}_R\mathbf{Mod}$ that is not enjoyed by the category **Sets**, for example: Morphisms can be added.

In Corollary 7.17, we gave another description of direct sum in terms of arrows:

There is a map $\rho \colon M \to S$ with $\rho s = s$; moreover, $\ker \rho = \operatorname{im} j$, $\operatorname{im} \rho = \operatorname{im} i$, and $\rho(s) = s$ for every $s \in \operatorname{im} \rho$.

This description makes sense in **Sets**, but it does not make sense in arbitrary categories because the image of a morphism may fail to be defined. For example, the morphisms in $\mathcal{C}(G)$ [see Example 7.25(vi)] are elements in $\operatorname{Hom}(*, *) = G$, not functions, and so the image of a morphism has no obvious meaning.

However, we can define direct summand categorically: An object S is (equivalent to) a retract of an object M if there exist morphisms

$$i \colon S \to M \quad \text{and} \quad p \colon M \to S$$

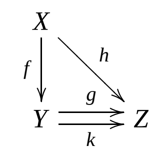
for which $pi = 1_S$ and $(ip)^2 = ip$ (for modules, define $\rho = ip$).

One of the nice aspects of thinking in a categorical way is that it enables us to see analogies that might not have been recognized before. For example, we shall soon see that direct sum in ${}_R\mathbf{Mod}$ is the same notion as disjoint union in **Sets**.

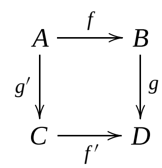
We begin with a very formal definition.

Definition. A ***diagram*** in a category $\mathcal{C}$ is a directed multigraph[9] whose vertices are objects in $\mathcal{C}$ and whose arrows are morphisms in $\mathcal{C}$.

For example,



is a diagram in a category, as is



If we think of an arrow as a "one-way street," then a *path* in a diagram is a "walk" from one vertex to another taking care never to walk the wrong way. A path in a diagram may be regarded as a composite of morphisms.

Definition. A diagram ***commutes*** if, for each pair of vertices A and B, any two paths from A to B are equal; that is, the composites are the same morphism.

For example, the triangular diagram above commutes if $gf = h$ and $kf = h$, and the square diagram above commutes if $gf = f'g'$. The term *commutes* in this context arises from this last example.

If A and B are subsets of a set S, then their intersection is defined:

$$A \cap B = \{s \in S : s \in A \text{ and } s \in B\}$$

(if two sets are not given as subsets, then their intersection may not be what one expects: for example, if $\mathbb{Q}$ is defined as all equivalence classes of ordered pairs (m, n) of integers with $n \neq 0$, then $\mathbb{Z} \cap \mathbb{Q} = \varnothing$).

We can force two overlapping subsets A and B to be disjoint by "disjointifying" them. Consider the cartesian product $(A \cup B) \times \{1, 2\}$, and consider the subsets $A' = A \times \{1\}$ and $B' = B \times \{2\}$. It is plain that $A' \cap B' = \varnothing$, for a point in the intersection would have coordinates $(a, 1) = (b, 2)$; this cannot be, for their second coordinates are not equal. We

[9] A ***directed multigraph*** consists of a set V, called ***vertices*** and, for each ordered pair $(u, v) \in V \times V$, a (possibly empty) set $\mathrm{arr}(u, v)$, called ***arrows*** from u to v.

call $A' \cup B'$ the ***disjoint union*** of A and B. Let us take note of the functions $\alpha\colon A \to A'$ and $\beta\colon B \to B'$, given by $\alpha\colon a \mapsto (a, 1)$ and $\beta\colon b \mapsto (b, 2)$. We denote the disjoint union $A' \cup B'$ by $A \sqcup B$.

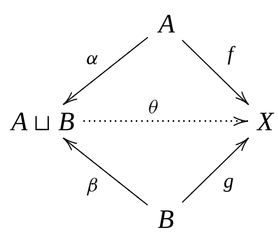
If there are functions $f\colon A \to X$ and $g\colon B \to X$, for some set X, then there is a unique function $h\colon A \sqcup B \to X$ given by

$$h(u) = \begin{cases} f(u) & \text{if } u \in A; \\ g(u) & \text{if } u \in B. \end{cases}$$

The function h is well-defined because A and B are disjoint.

Here is a way to describe this construction *categorically* (i.e., with diagrams).

Definition. If A and B are objects in a category $\mathcal{C}$, then their ***coproduct***, denoted by $A \sqcup B$, is an object C in $\text{obj}(\mathcal{C})$ together with ***injection morphisms*** $\alpha\colon A \to A \sqcup B$ and $\beta\colon B \to A \sqcup B$, such that, for every object X in $\mathcal{C}$ and every pair of morphisms $f\colon A \to X$ and $g\colon B \to X$, there exists a unique morphism $\theta\colon A \sqcup B \to X$ making the following diagram commute (i.e., $\theta\alpha = f$ and $\theta\beta = g$).



Here is the formal proof that the set $A \sqcup B = A' \cup B' \subseteq (A \cup B) \times \{1, 2\}$ just constructed is the coproduct in **Sets**. If X is any set and if $f\colon A \to X$ and $g\colon B \to X$ are any given functions, then there is a function $\theta\colon A \sqcup B \to X$ that extends both f and g. If $c \in A \sqcup B$, then either $c = (a, 1) \in A'$ or $c = (b, 2) \in B'$. Define $\theta((a, 1)) = f(a)$ and define $\theta((b, 2)) = g(b)$, so that $\theta\alpha = f$ and $\theta\beta = g$. Let us show that θ is the unique function on $A \sqcup B$ extending both f and g. If $\psi\colon A \sqcup B \to X$ satisfies $\psi\alpha = f$ and $\psi\beta = g$, then

$$\psi(\alpha(a)) = \psi((a, 1)) = f(a) = \theta((a, 1))$$

and, similarly,

$$\psi((b, 2)) = g(b).$$

Therefore, ψ agrees with θ on $A' \cup B' = A \sqcup B$, and so $\psi = \theta$.

We do not assert that coproducts always exist; in fact, it is easy to construct examples of categories in which a pair of objects does not have a coproduct (see Exercise 7.21 on page 458). Our argument, however, shows that coproducts do exist in **Sets**, where they are disjoint unions. Coproducts exist in the category of groups, and they are called ***free products***; free groups turn out to be free products of infinite cyclic groups (analogous to free abelian groups being direct sums of infinite cyclic groups). A theorem of A. G. Kurosh states that every subgroup of a free product is itself a free product.

Proposition 7.26. *If A and B are R-modules, then their coproduct in ${}_R\mathbf{Mod}$ exists, and it is the direct sum $C = A \sqcup B$.*

Proof. The statement of the proposition is not complete, for a coproduct requires injection morphisms α and β. The underlying set of $C = A \sqcup B$ is the cartesian product $A \times B$, and so we may define $\alpha\colon A \to C$ by $\alpha\colon a \mapsto (a, 0)$ and $\beta\colon B \to C$ by $\beta\colon b \mapsto (0, b)$.

Now let X be a module, and let $f\colon A \to X$ and $g\colon B \to X$ be homomorphisms. Define $\theta\colon C \to X$ by $\theta\colon (a, b) \mapsto f(a) + g(b)$. First, the diagram commutes: If $a \in A$, then $\theta\alpha(a) = \theta((a, 0)) = f(a)$ and, similarly, if $b \in B$, then $\theta\beta(b) = \theta((0, b)) = g(b)$. Finally, θ is unique. If $\psi\colon C \to X$ makes the diagram commute, then $\psi((a, 0)) = f(a)$ for all $a \in A$ and $\psi((0, b)) = g(b)$ for all $b \in B$. Since ψ is a homomorphism, we have

$$\begin{aligned}\psi((a, b)) &= \psi((a, 0) + (0, b)) \\ &= \psi((a, 0)) + \psi((0, b)) = f(a) + g(b).\end{aligned}$$

Therefore, $\psi = \theta$. •

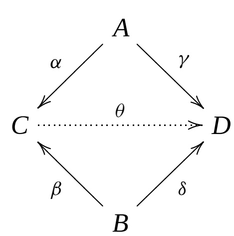
Let us give the explicit formula for the map θ in the proof of Proposition 7.26. If $f\colon A \to X$ and $g\colon B \to X$ are the given homomorphisms, then $\theta\colon A \oplus B \to X$ is given by

$$\theta\colon (a, b) \mapsto f(a) + g(b).$$

The outline of the proof of the next proposition will be used frequently; we have already seen it in our proof of Lemma 5.74, when we proved that the rank of a nonabelian free group is well-defined.

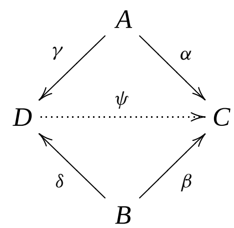
Proposition 7.27. *If $\mathcal{C}$ is a category and if A and B are objects in $\mathcal{C}$, then any two coproducts of A and B, should they exist, are equivalent.*

Proof. Suppose that C and D are coproducts of A and B. In more detail, assume that $\alpha\colon A \to C$, $\beta\colon B \to C$, $\gamma\colon A \to D$, and $\delta\colon B \to D$ are injection morphisms. If, in the defining diagram for C, we take $X = D$, then there is a morphism $\theta\colon C \to D$ making the diagram commute.

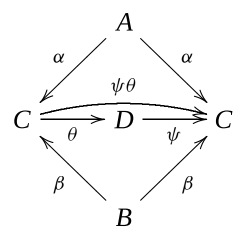


Similarly, if, in the defining diagram for D, we take $X = C$, we obtain a morphism

$\psi: D \to C$ making the diagram commute.



Consider now the following diagram, which arises from the juxtaposition of these two diagrams.

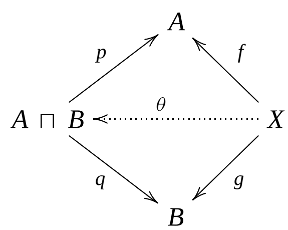


This diagram commutes because $\psi\theta\alpha = \psi\gamma = \alpha$ and $\psi\theta\beta = \psi\delta = \beta$. But plainly, the identity morphism $1_C: C \to C$ also makes this diagram commute. By the uniqueness of the dotted arrow in the defining diagram for coproduct, $\psi\theta = 1_C$. The same argument, mutatis mutandis, shows that $\theta\psi = 1_D$. We conclude that $\theta: C \to D$ is an equivalence. •

Informally, an object S in a category $\mathcal{C}$ is called a ***solution*** to a ***universal mapping problem*** if it is defined by a diagram such that, whenever we vary an object X and various morphisms in the diagram, there exists a unique morphism making the new diagram commute. The "metatheorem" is that solutions, if they exist, are unique to unique equivalence. The proof just given is the prototype for proving the metatheorem (if we wax categorical, then the statement of the metatheorem can be made precise, and we can then prove it; see Exercise 7.29 on page 459 for an illustration, and see Mac Lane, *Categories for the Working Mathematician*, Chapter III, for appropriate definitions, statement, and proof). There are two steps. First, if C and D are solutions, get morphisms $\theta: C \to D$ and $\psi: D \to C$ by setting $X = D$ in the diagram showing that C is a solution, and by setting $X = C$ in the corresponding diagram showing that D is a solution. Second, set $X = C$ in the diagram for C and show that both $\psi\theta$ and 1_C are "dotted" morphisms making the diagram commute; as such a dotted morphism is unique, conclude that $\psi\theta = 1_C$. Similarly, the other composite $\theta\psi = 1_D$, and so θ is an equivalence.

Definition. If A and B are objects in a category $\mathcal{C}$, then their ***product***, denoted by $A \sqcap B$, is an object $P \in \mathcal{C}$ and morphisms $p: P \to A$ and $q: P \to B$, such that, for every object $X \in \mathcal{C}$ and every pair of morphisms $f: X \to A$ and $g: X \to B$, there exists a unique

morphism $\theta\colon X \to P$ making the following diagram commute:



The cartesian product $P = A \times B$ of two sets A and B is the categorical product in **Sets**. Define $p\colon A \times B \to A$ by $p\colon (a, b) \mapsto a$ and define $q\colon A \times B \to B$ by $q\colon (a, b) \mapsto b$. If X is a set and $f\colon X \to A$ and $g\colon X \to B$ are functions, then the reader may show that $\theta\colon X \to A \times B$, defined by $\theta\colon x \mapsto (f(x), g(x)) \in A \times B$, satisfies the necessary conditions.

Proposition 7.28. *If A and B are objects in a category $\mathcal{C}$, then any two products of A and B, should they exist, are equivalent.*

Proof. Adapt the proof of the prototype, Proposition 7.27 •

The reader should note that the defining diagram for product is obtained from the diagram for coproduct by reversing all the arrows. A similar reversal of arrows can be seen in Exercise 7.18 on page 441: The diagram characterizing a surjection in ${}_R\mathbf{Mod}$ is obtained by reversing all the arrows in the diagram that characterizes an injection. If S is a solution to a universal mapping problem posed by a diagram $\mathcal{D}$, let $\mathcal{D}'$ be the diagram obtained from $\mathcal{D}$ by reversing all its arrows. If S' is a solution to the universal mapping problem posed by $\mathcal{D}'$, then we call S and S' ***duals***. There are examples of categories in which an object and its dual object both exist, and there are examples in which an object exists but its dual does not.

What is the product of two modules?

Proposition 7.29. *If R is a commutative ring and A and B are R-modules, then their (categorical) product $A \sqcap B$ exists; in fact,*

$$A \sqcap B \cong A \sqcup B.$$

Remark. Thus, the product and coproduct of two objects, though distinct in **Sets**, coincide in ${}_R\mathbf{Mod}$. ◂

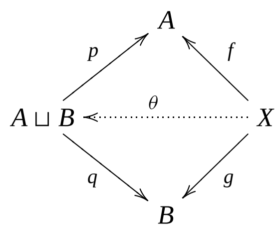
Proof. In Proposition 7.15(iii), we characterized $M \cong A \sqcup B$ by the existence of projection and injection morphisms

$$A \underset{p}{\overset{i}{\rightleftarrows}} M \underset{j}{\overset{q}{\rightleftarrows}} B$$

satisfying the equations

$$pi = 1_A,\ qj = 1_B,\ pj = 0,\ qi = 0 \quad \text{and} \quad ip + jq = 1_M.$$

If X is a module and $f : X \to A$ and $g : X \to B$ are homomorphisms, define $\theta : X \to A \sqcup B$ by $\theta(x) = if(x) + jg(x)$. The product diagram



commutes because, for all $x \in X$,

$$p\theta(x) = pif(x) + pjg(x) = pif(x) = f(x)$$

(using the given equations) and, similarly, $q\theta(x) = g(x)$. To prove uniqueness of θ, note that the equation $ip + jq = 1_{A \sqcup B}$ gives

$$\psi = ip\psi + jq\psi = if + jg = \theta. \quad \bullet$$

Exercise 7.23 on page 458 shows that direct products are products in **Groups**.

There are (at least) two ways to extend the notion of direct sum of modules from two summands to an indexed family of summands.

Definition. Let R be a commutative ring and let $\{A_i : i \in I\}$ be an indexed family of R-modules. The ***direct product*** $\prod_{i \in I} A_i$ is the cartesian product [i.e., the set of all I-tuples[10] (a_i) whose ith coordinate a_i lies in A_i for every i] with coordinatewise addition and scalar multiplication:

$$(a_i) + (b_i) = (a_i + b_i)$$
$$r(a_i) = (ra_i),$$

where $r \in R$ and $a_i, b_i \in A_i$ for all i.

The ***direct sum***, denoted by $\sum_{i \in I} A_i$ (and also by $\bigoplus_{i \in I} A_i$), is the submodule of $\prod_{i \in I} A_i$ consisting of all (a_i) having only finitely many nonzero coordinates.

Each $m \in \sum_{i \in I} A_i$ has a unique expression of the form

$$m = \sum_{i \in I} \alpha_i(a),$$

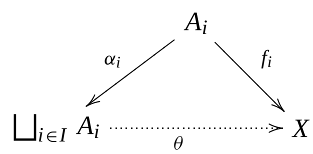
where $a_i \in A_i$. $\alpha_i(a)$ is the I-tuple in $\prod_i A_i$ whose ith coordinate is a_i and all other coordinates are 0, and ***almost all*** $a_i = 0$; that is, only finitely many a_i can be nonzero.

Note that if the index set I is finite, then $\prod_{i \in I} A_i = \sum_{i \in I} A_i$. On the other hand, when I is infinite and infinitely many $A_i \neq 0$, then the direct sum is a proper submodule of the direct product (moreover, in this case, they are almost never isomorphic).

We now extend the definitions of coproduct and product to a family of objects.

[10] An I-tuple is a function $f : I \to \bigcup_i A_i$ with $f(i) \in A_i$ for all $i \in I$.

Definition. Let $\mathcal{C}$ be a category, and let $\{A_i : i \in I\}$ be a family of objects in $\mathcal{C}$ indexed by a set I. A ***coproduct*** is an ordered pair $(C, \{\alpha_i : A_i \to C\})$, consisting of an object $C = \bigsqcup_{i \in I} A_i$ and a family $\{\alpha_i : A_i \to \bigsqcup_{i \in I} A_i$ for all $i \in I\}$ of ***injection*** morphisms, that satisfies the following property. For every object X equipped with morphisms $f_i : A_i \to X$, there exists a unique morphism $\theta : \bigsqcup_{i \in I} A_i \to X$ making the following diagram commute for each i:



As usual, coproducts are unique to equivalence should they exist.

We sketch the existence of the disjoint union of sets $\{A_i : i \in I\}$. First form the set $B = (\bigcup_{i \in I} A_i) \times I$, and then define

$$A_i' = \{(a_i, i) \in B : a_i \in A_i\}.$$

Then the disjoint union is $\bigsqcup_{i \in I} A_i = \bigcup_{i \in I} A_i'$ (of course, the disjoint union of two sets is a special case of this construction). The reader may show that $\bigsqcup_i A_i$ together with the functions $\alpha_i : A_i \to \bigsqcup_i A_i$ given by $\alpha_i : a_i \mapsto (a_i, i) \in \bigsqcup_i A_i$, comprise the coproduct in **Sets**; that is, we have described a solution to the universal mapping problem.

Proposition 7.30. *If $\{A_i : i \in I\}$ is a family of R-modules, then the direct sum $\sum_{i \in I} A_i$ is their coproduct in ${}_R$**Mod**.*

Proof. The statement of the proposition is not complete, for a coproduct requires injection morphisms α_i. Denote $\sum_{i \in I} A_i$ by C, and define $\alpha_i : A_i \to C$ by $a_i \mapsto \alpha_i(a)$ as follows: If $a_i \in A_i$, then $\alpha_i(a) \in C$ is the I-tuple whose ith coordinate is a_i and whose other coordinates are zero.

Now let X be a module and, for each $i \in I$, let $f_i : A_i \to X$ be homomorphisms. Define $\theta : C \to X$ by $\theta : (a_i) \mapsto \sum_i f_i(a_i)$ (note that this makes sense, for only finitely many a_i's are nonzero). First, the diagram commutes: If $a_i \in A_i$, then $\theta\alpha_i(a_i) = f_i(a_i)$. Finally, θ is unique. If $\psi : C \to X$ makes the diagram commute, then $\psi((a_i)) = f_i(a_i)$. Since ψ is a homomorphism, we have

$$\begin{aligned} \psi((a_i)) &= \psi\Big(\sum_i \alpha_i(a_i)\Big) \\ &= \sum_i \psi\alpha_i(a_i) = \sum_i f_i(a_i). \end{aligned}$$

Therefore, $\psi = \theta$. •

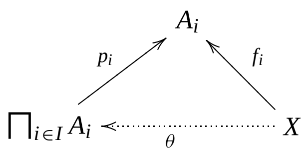
Let us make the formula for θ explicit. If $f_i : A_i \to X$ are given homomorphisms, then $\theta : \sum_{i \in I} A_i \to X$ is given by

$$\theta : (a_i) \mapsto \sum_{i \in I} f_i(a_i)$$

[of course, almost all the $a_i = 0$, so that there are only finitely many nonzero terms in the sum $\sum_{i\in I} f_i(a_i)$].

Here is the dual notion.

Definition. Let $\mathcal{C}$ be a category, and let $\{A_i : i \in I\}$ be a family of objects in $\mathcal{C}$ indexed by a set I. A ***product*** is an ordered pair $(C, \{p_i : C \to A_i\})$, consisting of an object $\prod_{i\in I} A_i$ and a family $\{p_i : C \to A_i$ for all $i \in I\}$ of ***projection*** morphisms, that satisfies the following condition. For every object X equipped with morphisms $f_i : X \to A_i$, there exists a unique morphism $\theta : X \to \prod_{i\in I} A_i$ making the following diagram commute for each i:



Products are unique to equivalence should they exist.

We let the reader prove that cartesian product is the product in **Sets**.

Proposition 7.31. *If $\{A_i : i \in I\}$ is a family of R-modules, then the direct product $C = \prod_{i\in I} A_i$ is their product in ${}_R$***Mod**.

Proof. The statement of the proposition is not complete, for a product requires projections. For each $j \in I$, define $p_j : C \to A_j$ by $p_j : (a_i) \mapsto a_j \in A_j$.

Now let X be a module and, for each $i \in I$, let $f_i : X \to A_i$ be a homomorphism. Define $\theta : X \to C$ by $\theta : x \mapsto (f_i(x))$. First, the diagram commutes: If $x \in X$, then $p_i\theta(x) = f_i(x)$. Finally, θ is unique. If $\psi : X \to C$ makes the diagram commute, then $p_i\psi(x) = f_i(a_i)$ for all i; that is, for each i, the ith coordinate of $\psi(x)$ is $f_i(x)$, which is also the ith coordinate of $\theta(x)$. Therefore, $\psi(x) = \theta(x)$ for all $x \in X$, and so $\psi = \theta$. •

The categorical viewpoint makes the next two proofs straightforward.

Theorem 7.32. *Let R be a commutative ring. For every R-module A and every family $\{B_i : i \in I\}$ of R-modules,*

$$\operatorname{Hom}_R\Big(A, \prod_{i\in I} B_i\Big) \cong \prod_{i\in I} \operatorname{Hom}_R(A, B_i),$$

via the R-isomorphism

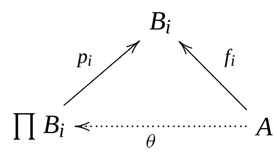
$$\varphi : f \mapsto (p_i f),$$

where the p_i are the projections of the product $\prod_{i\in I} B_i$.

Proof. It is easy to see that φ is additive. To see that φ is an R-map, note, for each i and each $r \in R$, that $p_i r f = r p_i f$; therefore,

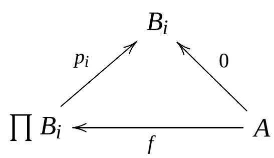
$$\varphi : rf \mapsto (p_i r f) = (r p_i f) = r(p_i f) = r\varphi(f).$$

Let us see that φ is surjective. If $(f_i) \in \prod \operatorname{Hom}_R(A, B_i)$, then $f_i\colon A \to B_i$ for every i.



By Proposition 7.31, $\prod B_i$ is the product in ${}_R\mathbf{Mod}$, and so there is a unique R-map $\theta\colon A \to \prod B_i$ with $p_i\theta = f_i$ for all i. Thus, $(f_i) = \varphi(\theta)$ and φ is surjective.

To see that φ is injective, suppose that $f \in \ker\varphi$; that is, $0 = \varphi(f) = (p_i f)$. Thus, $p_i f = 0$ for every i. Hence, the following diagram containing f commutes:



But the zero homomorphism also makes this diagram commute, and so the uniqueness of the arrow $A \to \prod B_i$ gives $f = 0$. •

Theorem 7.33. *For every R-module B and every family $\{A_i : i \in I\}$ of R-modules,*

$$\operatorname{Hom}_R\Big(\sum_{i\in I} A_i, B\Big) \cong \prod_{i\in I} \operatorname{Hom}_R(A_i, B),$$

via the R-isomorphism

$$f \mapsto (f\alpha_i),$$

where the α_i are the injections of the sum $\sum_{i\in I} A_i$.

Proof. This proof is similar to that of Theorem 7.32, and it is left to the reader. •

There are examples showing that $\operatorname{Hom}_R(A, \sum_i B_i) \ncong \sum_i \operatorname{Hom}_R(A, B_i)$ and that $\operatorname{Hom}_R(\prod_i A_i, B) \ncong \prod_i \operatorname{Hom}_R(A_i, B)$.

Corollary 7.34. *If A, A', B, and B' are R-modules. then there are isomorphisms*

$$\operatorname{Hom}_R(A, B \sqcup B') \cong \operatorname{Hom}_R(A, B) \sqcup \operatorname{Hom}_R(A, B')$$

and

$$\operatorname{Hom}_R(A \sqcup A', B) \cong \operatorname{Hom}_R(A, B) \sqcup \operatorname{Hom}_R(A', B).$$

Proof. When the index set is finite, the direct sum and the direct product of modules are equal. •

Example 7.35.
(i) In Example 7.6, we defined the *dual space* V^* of a vector space V over a field k to be the vector space of all its linear functionals:

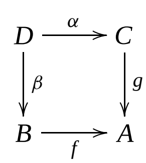
$$V^* = \mathrm{Hom}_k(V, k).$$

If $\dim(V) = n < \infty$, then Example 5.6 shows that $V = V_1 \oplus \cdots \oplus V_n$, where each V_i is one-dimensional. By Corollary 7.34, $V^* \cong \sum_i \mathrm{Hom}_k(V_i, k)$ is a direct sum of n one-dimensional spaces [for Exercise 7.5 on page 440 gives $\mathrm{Hom}_k(k, k) \cong k$], and so Exercise 7.26 on page 458 gives $\dim(V^*) = \dim(V) = n$. Thus, a finite-dimensional vector space and its dual space are isomorphic. It follows that the double dual, V^{**}, defined as $(V^*)^*$, is isomorphic to V when V is finite-dimensional.

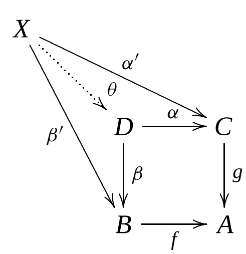
(ii) There are variations of dual spaces. In functional analysis, one encounters topological real vector spaces V, so that it makes sense to speak of *continuous* linear functionals. The *topological dual* V^* consists of all the continuous linear functionals, and it is important to know whether a space V is *reflexive*; that is, whether the analog of the isomorphism $V \to V^{**}$ for finite-dimensional spaces is a homeomorphism for these spaces. For example, that *Hilbert space* is reflexive is one of its important properties. ◀

We now present two dual constructions that are often useful.

Definition. Given two morphisms $f : B \to A$ and $g : C \to A$ in a category $\mathcal{C}$, a ***solution*** is an ordered triple (D, α, β) making the following diagram commute:



A ***pullback*** (or *fibered product*) is a solution (D, α, β) that is "best" in the following sense: For every solution (X, α', β'), there exists a unique morphism $\theta : X \to D$ making the following diagram commute:



Pullbacks, when they exist, are unique to equivalence; the proof is in the same style as the proof that coproducts are unique.

Proposition 7.36. *The pullback of two maps $f\colon B \to A$ and $g\colon C \to A$ in ${}_R$**Mod** exists.*

Proof. Define

$$D = \{(b, c) \in B \sqcup C : f(b) = g(c)\},$$

define $\alpha\colon D \to C$ to be the restriction of the projection $(b, c) \mapsto c$, and define $\beta\colon D \to B$ to be the restriction of the projection $(b, c) \mapsto b$. It is easy to see that (D, α, β) is a solution.

If (X, α', β') is another solution, define a map $\theta\colon X \to D$ by $\theta\colon x \mapsto (\beta'(x), \alpha'(x))$. The values of θ do lie in D, for $f\beta'(x) = g\alpha'(x)$ because X is a solution. We let the reader prove that the diagram commutes and that θ is unique. •

Example 7.37.
(i) That B and C are subsets of a set A can be restated as saying that there are inclusion maps $i : B \to A$ and $j : C \to A$. The reader will enjoy proving that the pullback D exists in **Sets**, and that $D = B \cap C$.

(ii) Pullbacks exist in **Groups**: They are certain subgroups of a direct product constructed as in the proof of Proposition 7.36.

(iii) If $f : B \to A$ is a homomorphism, then ker f is the pullback of the following diagram:

$$\begin{array}{ccc} & & 0 \\ & & \downarrow \\ B & \xrightarrow[f]{} & A \end{array}$$

The pullback is $\{(b, 0) \in B \sqcup \{0\} : fb = 0\} \cong \ker f$. ◂

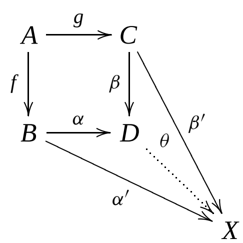
Here is the dual construction.

Definition. Given two morphisms $f\colon A \to B$ and $g\colon A \to C$ in a category $\mathcal{C}$, a ***solution*** is an ordered triple (D, α, β) making the following diagram commute:

$$\begin{array}{ccc} A & \xrightarrow{g} & C \\ {\scriptstyle f}\downarrow & & \downarrow{\scriptstyle \beta} \\ B & \xrightarrow{\alpha} & D \end{array}$$

A ***pushout*** (or *fibered sum*) is a solution (D, α, β) that is "best" in the following sense: for every solution (X, α', β'), there exists a unique morphism $\theta\colon D \to X$ making the

following diagram commute:



Again, pushouts are unique to equivalence when they exist.

Proposition 7.38. *The pushout of two maps $f : A \to B$ and $g : A \to C$ in ${}_R$**Mod** exists.*

Proof. It is easy to see that

$$S = \{\big(f(a), -g(a)\big) \in B \sqcup C : a \in A\}$$

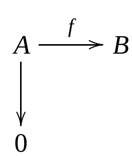
is a submodule of $B \sqcup C$. Define $D = (B \sqcup C)/S$, define $\alpha : B \to D$ by $b \mapsto (b, 0) + S$, and define $\beta : C \to D$ by $c \mapsto (0, c) + S$. It is easy to see that (D, α, β) is a solution.

Given another solution (X, α', β'), define the map $\theta : D \to X$ by $\theta : (b, c) + S \mapsto \alpha'(b) + \beta'(c)$. Again, we let the reader prove commutativity of the diagram and uniqueness of θ. •

Pushouts in **Groups** are quite interesting; for example, the pushout of two injective homomorphisms is called a *free product with amalgamation*.

Example 7.39.
(i) If B and C are subsets of a set A, then there are inclusion maps $i : B \cap C \to B$ and $j : B \cap C \to B$. The reader will enjoy proving that the pushout D exists in **Sets**, and that D is their union $B \cup C$.

(ii) If $f : A \to B$ is a homomorphism, then coker f is the pushout of the following diagram:



After all, the pushout here is the quotient $(\{0\} \sqcup B)/S$, where $S = \{(0, fa)\}$, and so $(\{0\} \sqcup B)/S \cong B/\operatorname{im} f = \operatorname{coker} f$. ◂

EXERCISES

7.20 (i) Prove, in every category $\mathcal{C}$, that each object $A \in \mathcal{C}$ has a unique identity morphism.

(ii) If f is an equivalence in a category, prove that its inverse is unique.

7.21 (i) Let X be a partially ordered set, and let $a, b \in X$. Show, in $\mathbf{PO}(X)$ [defined in Example 7.25(v)], that the coproduct $a \sqcup b$ is the least upper bound of a and b, and that the product $a \sqcap b$ is the greatest lower bound.

(ii) Let Y be a set, and let $\mathcal{P}(Y)$ denote its *power set*; that is, $\mathcal{P}(Y)$ is the family of all the subsets of Y. Now regard $\mathcal{P}(Y)$ as a partially ordered set under inclusion. If A and B are subsets of Y, show, in $\mathbf{PO}(\mathcal{P}(Y))$, that the coproduct $A \sqcup B = A \cup B$ and that the product $A \sqcap B = A \cap B$.

(iii) Give an example of a category in which there are two objects whose coproduct does not exist.

Hint. See Exercise 6.43 on page 374.

7.22 Prove that **Groups** is not a pre-additive category.
Hint. If G is not abelian and $f, g \colon G \to G$ are homomorphisms, show that the function $x \mapsto f(x)g(x)$ may not be a homomorphism.

7.23 If A and B are (not necessarily abelian) groups, prove that $A \sqcap B = A \times B$ (direct product) in **Groups**.

7.24 If G is a finite abelian group, prove that $\text{Hom}_{\mathbb{Z}}(\mathbb{Q}, G) = 0$.

7.25 Let $\{M_i : i \in I\}$ be a family of modules and, for each i, let N_i be a submodule of M_i. Prove that

$$\Big(\sum_i M_i\Big)/\Big(\sum_i N_i\Big) \cong \sum_i \big(M_i/N_i\big).$$

7.26 (i) Let $v_1, \ldots, v_n$ be a basis of a vector space V over a field k, so that every $v \in V$ has a unique expression

$$v = a_1 v_1 + \cdots + a_n v_n,$$

where $a_i \in k$ for $i = 1, \ldots, n$. For each i, prove that the function $v_i^* \colon V \to k$, defined by $v_i^* \colon v \mapsto a_i$, lies in the dual space V^*.

(ii) Prove that $v_1^*, \ldots, v_n^*$ is a linearly independent list in V^*.

(iii) Use Example 7.35(i) to conclude that $v_1^*, \ldots, v_n^*$ is a basis of V^* (it is called the ***dual basis*** of $v_1, \ldots, v_n$).

(iv) If $f \colon V \to V$ is a linear transformation, let A be the matrix of f with respect to a basis $v_1, \ldots, v_n$ of V; that is, the ith column of A consists of the coordinates of $f(v_i)$ in terms of the given basis $v_1, \ldots, v_n$. Prove that the matrix of the induced map $f^* \colon V^* \to V^*$ with respect to the dual basis is A^t, the transpose of A.

7.27 Given a map $\sigma \colon \prod B_i \to \prod C_j$, find a map $\widetilde{\sigma}$ making the following diagram commute,

$$\begin{array}{ccc}
\text{Hom}(A, \prod B_i) & \xrightarrow{\ \sigma\ } & \text{Hom}(A, \prod C_j) \\
\downarrow{\scriptstyle \tau} & & \downarrow{\scriptstyle \tau'} \\
\prod \text{Hom}(A, B_i) & \xrightarrow{\ \widetilde{\sigma}\ } & \prod \text{Hom}(A, C_j),
\end{array}$$

where τ and τ' are the isomorphisms of Theorem 7.32.
Hint. If $f \in \operatorname{Hom}(A, \prod B_i)$, define $\widetilde{\sigma} : (f_i) \mapsto (p_j \sigma f)$; that is, the jth coordinate of $\widetilde{\sigma}(f_i))$ is the jth coordinate of $\sigma(f) \in \prod C_j$.

7.28 (i) Given a pushout diagram in ${}_R\mathbf{Mod}$

$$\begin{array}{ccc} A & \xrightarrow{g} & C \\ {\scriptstyle f}\downarrow & & \downarrow{\scriptstyle \beta} \\ B & \xrightarrow[\alpha]{} & D \end{array}$$

prove that g injective implies α injective, and that g surjective implies α surjective. Thus, parallel arrows have the same properties.

(ii) Given a pullback diagram in ${}_R\mathbf{Mod}$

$$\begin{array}{ccc} D & \xrightarrow{\alpha} & C \\ {\scriptstyle \beta}\downarrow & & \downarrow{\scriptstyle g} \\ B & \xrightarrow[f]{} & A \end{array}$$

prove that f injective implies α injective, and that f surjective implies α surjective. Thus, parallel arrows have the same properties.

7.29 Definition. An object A in a category $\mathcal{C}$ is called an ***initial object*** if, for every object C in $\mathcal{C}$, there exists a unique morphism $A \to C$.

An object Ω in a category $\mathcal{C}$ is called a ***terminal object*** if, for every object C in $\mathcal{C}$, there exists a unique morphism $C \to \Omega$.

(i) Prove the uniqueness of initial and terminal objects, if they exist. Give an example of a category which contains no initial object. Give an example of a category that contains no terminal object.

(ii) If Ω is a terminal object in a category $\mathcal{C}$, prove, for any $G \in \operatorname{obj}(\mathcal{C})$, that the projections $\lambda : G \sqcap \Omega \to G$ and $\rho : \Omega \sqcap G \to G$ are equivalences.

(iii) Let A and B be objects in a category $\mathcal{C}$. Define a new category $\mathcal{C}'$ whose objects are diagrams

$$A \xrightarrow{\alpha} C \xleftarrow{\beta} B,$$

where C is an object in $\mathcal{C}$ and α and β are morphisms in $\mathcal{C}$. Define a morphism in $\mathcal{C}'$ to be a morphism θ in $\mathcal{C}$ that makes the following diagram commute:

$$\begin{array}{ccccc} A & \xrightarrow{\alpha} & C & \xleftarrow{\beta} & B \\ {\scriptstyle 1_A}\downarrow & & \downarrow{\scriptstyle \theta} & & \downarrow{\scriptstyle 1_B} \\ A & \xrightarrow[\alpha']{} & C' & \xleftarrow[\beta']{} & B \end{array}$$

There is an obvious candidate for composition. Prove that $\mathcal{C}'$ is a category.

(iv) Prove that an initial object in $\mathcal{C}'$ is a coproduct in $\mathcal{C}$.

(v) Give an analogous construction showing that product is a terminal object in a suitable category.

7.30 A ***zero object*** in a category $\mathcal{C}$ is an object Z that is both an initial object and a terminal object.

(i) Prove that $\{0\}$ is a zero object in ${}_R\mathbf{Mod}$.

(ii) Prove that $\varnothing$ is an initial object in **Sets**.

(iii) Prove that any one-point set is a terminal object in **Sets**.

(iv) Prove that a zero object does not exist in **Sets**.

7.31 (i) Assuming that coproducts exist, prove associativity:

$$A \sqcup (B \sqcup C) \cong (A \sqcup B) \sqcup C.$$

(ii) Assuming that products exist, prove associativity:

$$A \sqcap (B \sqcap C) \cong (A \sqcap B) \sqcap C.$$

7.32 Let C_1, C_2, D_1, D_2 be objects in a category $\mathcal{C}$.

(i) If there are morphisms $f_i \colon C_i \to D_i$, for $i = 1, 2$, and if $C_1 \sqcap C_2$ and $D_1 \sqcap D_2$ exist, prove that there exists a unique morphism $f_1 \sqcap f_2$ making the following diagram commute:

$$\begin{array}{ccc} C_1 \sqcap C_2 & \xrightarrow{f_1 \sqcap f_2} & D_1 \sqcap D_2 \\ {\scriptstyle p_i}\downarrow & & \downarrow{\scriptstyle q_i} \\ C_i & \xrightarrow[f_i]{} & D_i, \end{array}$$

where p_i and q_i are projections.

(ii) If there are morphisms $g_i : X \to C_i$, where X is an object in $\mathcal{C}$ and $i = 1, 2$, prove that there is a unique morphism (g_1, g_2) making the following diagram commute:

$$\begin{array}{ccccc} & {\scriptstyle g_1}\swarrow & X & \searrow{\scriptstyle g_2} & \\ & & \downarrow{\scriptstyle (g_1, g_2)} & & \\ C_1 & \xleftarrow[p_1]{} & C_1 \sqcap C_2 & \xrightarrow[p_2]{} & C_2, \end{array}$$

where the p_i are projections.

Hint. First define an analog of the diagonal $\Delta_X : X \to X \times X$ in **Sets**, given by $x \mapsto (x, x)$, and then define $(g_1, g_2) = (g_1 \sqcap g_2)\Delta_X$.

7.33 Let $\mathcal{C}$ be a category having finite products and a terminal object Ω. A ***group object*** in $\mathcal{C}$ is a quadruple (G, μ, η, ϵ), where G is an object in $\mathcal{C}$, $\mu \colon G \prod G \to G$, $\eta \colon G \to G$, and $\epsilon \colon \Omega \to G$ are morphisms, so that the following diagrams commute:

Associativity:

$$\begin{array}{ccc} G \sqcap G \sqcap G & \xrightarrow{1 \sqcap \mu} & G \sqcap G \\ {\scriptstyle \mu \sqcap 1}\downarrow & & \downarrow{\scriptstyle \mu} \\ G \sqcap G & \xrightarrow[\mu]{} & G \end{array}$$

Identity:

$$\begin{array}{ccccc} G \sqcap \Omega & \xrightarrow{1 \sqcap \epsilon} & G \sqcap G & \xleftarrow{\epsilon \sqcap 1} & \Omega \sqcap G \\ & {\scriptstyle\lambda}\searrow & \downarrow{\scriptstyle\mu} & \swarrow{\scriptstyle\rho} & \\ & & G & & \end{array}$$

where λ and ρ are the equivalences in Exercise 7.29(ii).
Inverse:

$$\begin{array}{ccccc} G & \xrightarrow{(1,\eta)} & G \sqcap G & \xleftarrow{(\eta,1)} & G \\ {\scriptstyle\omega}\downarrow & & \downarrow{\scriptstyle\mu} & & \downarrow{\scriptstyle\omega} \\ \Omega & \xrightarrow[\epsilon]{} & G & \xleftarrow[\epsilon]{} & \Omega \end{array}$$

where $\omega\colon G \to \Omega$ is the unique morphism to the terminal object.

(i) Prove that a group object in **Sets** is a group.

(ii) Prove that a group object in **Groups** is an abelian group.
Hint. Use Exercise 2.73 on page 95.

7.3 FUNCTORS

Functors[11] are homomorphisms of categories.

Definition. Recall that obj($\mathcal{C}$) denotes the class of all the objects in a category $\mathcal{C}$. If $\mathcal{C}$ and $\mathcal{D}$ are categories, then a ***functor*** $T : \mathcal{C} \to \mathcal{D}$ is a function such that

(i) if $A \in \text{obj}(\mathcal{C})$, then $T(A) \in \text{obj}(\mathcal{D})$;

(ii) if $f : A \to A'$ in $\mathcal{C}$, then $T(f) \colon T(A) \to T(A')$ in $\mathcal{D}$;

(iii) if $A \xrightarrow{f} A' \xrightarrow{g} A''$ in $\mathcal{C}$, then $T(A) \xrightarrow{T(f)} T(A') \xrightarrow{T(g)} T(A'')$ in $\mathcal{D}$ and

$$T(gf) = T(g)T(f);$$

(iv) for every $A \in \text{obj}(\mathcal{C})$,

$$T(1_A) = 1_{T(A)}.$$

Example 7.40.
(i) If $\mathcal{C}$ is a category, then the ***identity functor*** $1_{\mathcal{C}} : \mathcal{C} \to \mathcal{C}$ is defined by

$$1_{\mathcal{C}}(A) = A \text{ for all objects } A,$$

and

$$1_{\mathcal{C}}(f) = f \text{ for all morphisms } f.$$

[11]The term *functor* was coined by the philosopher R. Carnap, and S. Mac Lane thought it was the appropriate term in this context.

(ii) If $\mathcal{C}$ is a category and $A \in \text{obj}(\mathcal{C})$, then the **Hom *functor*** $T_A \colon \mathcal{C} \to$ **Sets** is defined by

$$T_A(B) = \text{Hom}(A, B) \text{ for all } B \in \text{obj}(\mathcal{C}),$$

and if $f \colon B \to B'$ in $\mathcal{C}$, then $T_A(f) \colon \text{Hom}(A, B) \to \text{Hom}(A, B')$ is given by

$$T_A(f) \colon h \mapsto fh.$$

We call $T_A(f)$ the ***induced map***, and we denote it by

$$T_A(f) = f_* \colon h \mapsto fh.$$

Because of the importance of this example, we will verify the parts of the definition in detail. First, the very definition of category says that $\text{Hom}(A, B)$ is a set. Note that the composite fh makes sense:

$$A \xrightarrow{h} B \xrightarrow{f} B' \quad (\text{composite } A \xrightarrow{fh} B').$$

Suppose now that $g \colon B \to B''$. Let us compare the functions

$$(gf)_*, g_* f_* \colon \text{Hom}(A, B) \to \text{Hom}(A, B'').$$

If $h \in \text{Hom}(A, B)$, i.e., if $h \colon A \to B$, then

$$(gf)_* \colon h \mapsto (gf)h;$$

on the other hand,

$$g_* f_* \colon h \mapsto fh \mapsto g(fh),$$

as desired. Finally, if f is the identity map $1_A \colon A \to A$, then

$$(1_A)_* \colon h \mapsto 1_A h = h$$

for all $h \in \text{Hom}(A, B)$, so that $(1_A)_* = 1_{\text{Hom}(A,B)}$.

If we denote $\text{Hom}(A, \ )$ by T_A, then Theorem 7.32 says that T_A preserves products: $T_A\big(\prod_i B_i\big) \cong \prod_i T_A(B_i)$.

(iii) If R is a commutative ring and A is an R-module, then the Hom functor $T_A \colon {}_R\mathbf{Mod} \to$ **Sets** has more structure. We have seen, in Proposition 7.5, that $\text{Hom}_R(A, B)$ is an R-module; we now show that if $f \colon B \to B'$, then the induced map $f_* \colon \text{Hom}_R(A, B) \to \text{Hom}_R(A, B')$, given by $h \mapsto fh$, is an R-map. First, f_* is additive: If $h, h' \in \text{Hom}(A, B)$, then for all $a \in A$,

$$\begin{aligned} f_*(h + h') = f(h + h') \colon a &\mapsto f(ha + h'a) \\ &= fha + fh'a = (f_*(h) + f_*(h'))(a), \end{aligned}$$

so that $f_*(h + h') = f_*(h) + f_*(h')$. Second, f_* preserves scalars. Recall that if $r \in R$ and $h \in \text{Hom}(A, B)$, then $rh \colon a \mapsto h(ra)$. Thus,

$$f_*(rh) \colon a \mapsto f(rh)(a) = fh(ra),$$

while

$$rf_*(h) = rfh \colon a \mapsto fh(ra).$$

Therefore, $f_*(rh) = (rf)_*(h)$.

In particular, if R is a field, then the Hom_R's are vector spaces and the induced maps are linear transformations.

(iv) Let $\mathcal{C}$ be a category, and let $A \in \text{obj}(\mathcal{C})$. Define $T : \mathcal{C} \to \mathcal{C}$ by $T(C) = A$ for every $C \in \text{obj}(\mathcal{C})$, and $T(f) = 1_A$ for every morphism f in $\mathcal{C}$. Then T is a functor, called the ***constant functor*** at A.

(v) If $\mathcal{C} = \textbf{Groups}$, define the ***forgetful functor*** $U \colon \textbf{Groups} \to \textbf{Sets}$ by $U(G)$ is the "underlying" set of a group G and $U(f)$ regards a homomorphism f as a mere function. Strictly speaking, a group is an ordered pair (G, μ), where G is its (underlying) set and $\mu \colon G \times G \to G$ is its operation, and $U((G, \mu)) = G$; the functor U "forgets" the operation and remembers only the set.

There are many variants. For example, an R-module is an ordered triple (M, α, σ), where M is a set, $\alpha \colon M \times M \to M$ is addition, and $\sigma : R \times M \to M$ is scalar multiplication. There are forgetful functors $U' \colon {}_R\textbf{Mod} \to \textbf{Ab}$ with $U'((M, \alpha, \sigma)) = (M, \alpha)$, and $U'' \colon {}_R\textbf{Mod} \to \textbf{Sets}$ with $U''(M, \alpha, \sigma)) = M$, for example. ◂

The following result is useful, even though it is very easy to prove.

Proposition 7.41. *If $T : \mathcal{C} \to \mathcal{D}$ is a functor, and if $f : A \to B$ is an equivalence in $\mathcal{C}$, then $T(f)$ is an equivalence in $\mathcal{D}$.*

Proof. If g is the inverse of f, apply T to the equations

$$gf = 1_A \text{ and } fg = 1_B. \quad \bullet$$

This proposition illustrates, admittedly at a low level, the reason why it is useful to give categorical definitions: Functors can recognize definitions phrased solely in terms of objects, morphisms, and diagrams. How could we prove this result in **Ab** if we regard an isomorphism as a homomorphism that is an injection and a surjection?

There is a second type of functor that reverses the direction of arrows.

Definition. If $\mathcal{C}$ and $\mathcal{D}$ are categories, then a ***contravariant functor*** $T : \mathcal{C} \to \mathcal{D}$ is a function such that

(i) if $C \in \text{obj}(\mathcal{C})$, then $T(C) \in \text{obj}(\mathcal{D})$;

(ii) if $f : C \to C'$ in $\mathcal{C}$, then $T(f) \colon T(C') \to T(C)$ in $\mathcal{D}$;

(iii) if $C \xrightarrow{f} C' \xrightarrow{g} C''$ in $\mathcal{C}$, then $T(C'') \xrightarrow{T(g)} T(C') \xrightarrow{T(f)} T(C)$ in $\mathcal{D}$ and

$$T(gf) = T(f)T(g);$$

(iv) for every $A \in \text{obj}(\mathcal{C})$,

$$T(1_A) = 1_{T(A)}.$$

To distinguish them from contravariant functors, the functors defined earlier are called ***covariant functors***.

Example 7.42.
(i) If $\mathcal{C}$ is a category and $B \in \text{obj}(\mathcal{C})$, then the ***contravariant* Hom *functor*** $T^B \colon \mathcal{C} \to \mathbf{Sets}$ is defined, for all $C \in \text{obj}(\mathcal{C})$, by

$$T^B(C) = \text{Hom}(C, B)$$

and if $f \colon C \to C'$ in $\mathcal{C}$, then $T^B(f) \colon \text{Hom}(C', B) \to \text{Hom}(C, B)$ is given by

$$T^B(f) \colon h \mapsto hf.$$

We call $T^B(f)$ the ***induced map***, and we denote it by

$$T^B(f) = f^* \colon h \mapsto hf.$$

Because of the importance of this example, we verify the axioms, showing that T^B is a (contravariant) functor. Note that the composite hf makes sense:

$$C \xrightarrow{f} C' \xrightarrow{h} B, \qquad hf \colon C \to B.$$

Given homomorphisms

$$C \xrightarrow{f} C' \xrightarrow{g} C'',$$

let us compare the functions

$$(gf)^*,\ f^*g^* \colon \text{Hom}(C'', B) \to \text{Hom}(C, B).$$

If $h \in \text{Hom}(C'', B)$ (i.e., if $h \colon C'' \to B$), then

$$(gf)^* \colon h \mapsto h(gf);$$

on the other hand,

$$f^*g^* \colon h \mapsto hg \mapsto (hg)f,$$

as desired. Finally, if f is the identity map $1_C \colon C \to C$, then

$$(1_C)^* \colon h \mapsto h1_C = h$$

for all $h \in \text{Hom}(C, B)$, so that $(1_C)^* = 1_{\text{Hom}(C,B)}$.

If $\text{Hom}(\ \ , B)$ is denoted by T^B, then Theorem 7.33 says that the contravariant functor T^B converts sums to products: $T^B\big(\sum_i A_i\big) \cong \prod_i T^B(A_i)$.

(ii) If R is a commutative ring and C is an R-module, then the contravariant Hom functor ${}_R\mathbf{Mod} \to \mathbf{Sets}$ has more structure. We show that if $f: C \to C'$ is an R-map, then the induced map $f^*\colon \text{Hom}_R(C', B) \to \text{Hom}_R(C, B)$, given by $h \mapsto hf$, is an R-map between R-modules. First, f^* is additive: If $g, h \in \text{Hom}(C', B)$, then for all $c' \in C'$,

$$\begin{aligned} f^*(g+h) = (g+h)f : c' &\mapsto (g+h)f(c') \\ &= gfc' + hfc' = (f^*(g) + f^*(h))(c'), \end{aligned}$$

so that $f^*(g+h) = f^*(g) + f^*(h)$. Second, f^* preserves scalars. Recall that if $r \in R$ and $h \in \text{Hom}(A, B)$, then $rh\colon a \mapsto h(ra)$. Thus,

$$f^*(rh)\colon c' \mapsto (rh)f(c') = h(rf(c')),$$

while

$$rf^*(h) = r(hf)\colon c' \mapsto hf(rc').$$

These are the same, because $rf(c') = f(rc')$, and so $f^*(rh) = rf^*(h)$.

In particular, if R is a field, then the Hom_R's are vector spaces and the induced maps are linear transformations. A special case of this is the ***dual space functor*** $\text{Hom}_k(\ \ , k)$, where k is a field. ◂

It is easy to see, as in Proposition 7.41, that every contravariant functor preserves equivalences; that is, if $T : \mathcal{C} \to \mathcal{D}$ is a contravariant functor, and if $f : C \to C'$ is an equivalence in $\mathcal{C}$, then $T(f)$ is an equivalence in $\mathcal{D}$.

Definition. If $\mathcal{C}$ and $\mathcal{D}$ are pre-additive categories, then a functor $T : \mathcal{C} \to \mathcal{D}$, of either variance, is called an ***additive functor*** if, for every pair of morphisms $f, g\colon A \to B$, we have

$$T(f+g) = T(f) + T(g).$$

It is easy to see that Hom functors ${}_R\mathbf{Mod} \to \mathbf{Ab}$ of either variance are additive functors.

Every covariant functor $T : \mathcal{C} \to \mathcal{D}$ gives rise to functions

$$T_{AB}\colon \text{Hom}(A, B) \to \text{Hom}(TA, TB),$$

for every A and B, defined by $h \mapsto T(h)$. If T is an additive functor between pre-additive categories, then each T_{AB} is a homomorphism of abelian groups; the analogous statement for contravariant functors is also true.

Here is a modest generalization of Corollary 7.34.

Proposition 7.43. *If $T : {}_R\mathbf{Mod} \to \mathbf{Ab}$ is an additive functor of either variance, then T preserves finite direct sums:*

$$T(A_1 \oplus \cdots \oplus A_n) \cong T(A_1) \oplus \cdots \oplus T(A_n).$$

Proof. By induction, it sufffices to prove that $T(A \oplus B) \cong T(A) \oplus T(B)$. Proposition 7.15(iii) characterizes $M = A \oplus B$ by maps $p : M \to A$, $q : M \to B$, $i : A \to M$, and $j : B \to M$ such that

$$pi = 1_A, \ qj = 1_B, \ pj = 0, \ qi = 0 \quad \text{and} \quad ip + jq = 1_M.$$

Since T is an additive functor, Exercise 7.34 on page 470 gives $T(0) = 0$, and so T preserves these equations. •

We have just seen that additive functors $T : {}_R\mathbf{Mod} \to \mathbf{Ab}$ preserve the direct sum of two modules:

$$T(A \oplus C) = T(A) \oplus T(C).$$

If we regard such a direct sum as a split short exact sequence, then we may rephrase this by saying that if

$$0 \to A \xrightarrow{i} B \xrightarrow{p} C \to 0$$

is a split short exact sequence, then so is

$$0 \to T(A) \xrightarrow{T(i)} T(B) \xrightarrow{T(p)} T(C) \to 0.$$

This leads us to the more general question: If

$$0 \to A \xrightarrow{i} B \xrightarrow{p} C \to 0$$

is any short exact sequence, not necessarily split, is

$$0 \to T(A) \xrightarrow{T(i)} T(B) \xrightarrow{T(p)} T(C) \to 0$$

also an exact sequence? Here is the answer for Hom functors (there is no misprint in the statement of the theorem: "$\to 0$" should not appear at the end of the sequences, and we shall discuss this point after the proof).

Theorem 7.44. *If*

$$0 \to A \xrightarrow{i} B \xrightarrow{p} C$$

is an exact sequence of R-modules, and if X is an R-module, then there is an exact sequence

$$0 \to \mathrm{Hom}_R(X, A) \xrightarrow{i_*} \mathrm{Hom}_R(X, B) \xrightarrow{p_*} \mathrm{Hom}_R(X, C).$$

Proof. (i) $\ker i_* = \{0\}$:

If $f \in \ker i_*$, then $f\colon X \to A$ and $i_*(f) = 0$; that is,

$$if(x) = 0 \text{ for all } x \in X.$$

Since i is injective, $f(x) = 0$ for all $x \in X$, and so $f = 0$.

(ii) $\operatorname{im} i_* \subseteq \ker p_*$:

If $g \in \operatorname{im} i_*$, then $g\colon X \to B$ and $g = i_*(f) = if$ for some $f\colon X \to A$. But $p_*(g) = pg = pif = 0$ because exactness of the original sequence, namely, $\operatorname{im} i = \ker p$, implies $pi = 0$.

(iii) $\ker p_* \subseteq \operatorname{im} i_*$:

If $g \in \ker p_*$, then $g\colon X \to B$ and $p_*(g) = pg = 0$. Hence, $pg(x) = 0$ for all $x \in X$, so that $g(x) \in \ker p = \operatorname{im} i$. Thus, $g(x) = i(a)$ for some $a \in A$; since i is injective, this element a is unique. Hence, the function $f\colon X \to A$, given by $f(x) = a$ if $g(x) = i(a)$, is well-defined. It is easy to check that $f \in \operatorname{Hom}_R(X, A)$; that is, f is an R-homomorphism. Since

$$g(x + x') = g(x) + g(x') = i(a) + i(a') = i(a + a'),$$

we have

$$f(x + x') = a + a' = f(x) + f(x').$$

A similar argument shows that $f(rx) = rf(x)$ for all $r \in R$. But, $i_*(f) = if$ and $if(x) = i(a) = g(x)$ for all $x \in X$; that is, $i_*(f) = g$, and so $g \in \operatorname{im} i_*$. •

Example 7.45.

Even if the map $p\colon B \to C$ in the original exact sequence is assumed to be surjective, the functored sequence need not end with "$\to 0$;" that is, $p_*\colon \operatorname{Hom}_R(X, B) \to \operatorname{Hom}_R(X, C)$ may fail to be surjective.

The abelian group $\mathbb{Q}/\mathbb{Z}$ consists of cosets $q + \mathbb{Z}$ for $q \in \mathbb{Q}$, and it easy to see that its element $\frac{1}{2} + \mathbb{Z}$ has order 2. It follows that $\operatorname{Hom}_{\mathbb{Z}}(\mathbb{I}_2, \mathbb{Q}/\mathbb{Z}) \neq \{0\}$, for it contains the nonzero homomorphism $[1] \mapsto \frac{1}{2} + \mathbb{Z}$.

Apply the functor $\operatorname{Hom}_{\mathbb{Z}}(\mathbb{I}_2, \ )$ to

$$0 \to \mathbb{Z} \xrightarrow{i} \mathbb{Q} \xrightarrow{p} \mathbb{Q}/\mathbb{Z} \to 0,$$

where i is the inclusion and p is the natural map. We have just seen that

$$\operatorname{Hom}_{\mathbb{Z}}(\mathbb{I}_2, \mathbb{Q}/\mathbb{Z}) \neq \{0\};$$

on the other hand, $\operatorname{Hom}_{\mathbb{Z}}(\mathbb{I}_2, \mathbb{Q}) = \{0\}$ because $\mathbb{Q}$ has no (nonzero) elements of finite order. Therefore, the induced map $p_*\colon \operatorname{Hom}_{\mathbb{Z}}(\mathbb{I}_2, \mathbb{Q}) \to \operatorname{Hom}_{\mathbb{Z}}(\mathbb{I}_2, \mathbb{Q}/\mathbb{Z})$ cannot be surjective. ◂

Definition. A covariant functor $T\colon {}_R\mathbf{Mod} \to \mathbf{Ab}$ is called ***left exact*** if exactness of

$$0 \to A \xrightarrow{i} B \xrightarrow{p} C$$

implies exactness of

$$0 \to T(A) \xrightarrow{T(i)} T(B) \xrightarrow{T(p)} T(C).$$

Thus, Theorem 7.44 shows that covariant Hom functors $\mathrm{Hom}_R(X,\ )$ are left exact functors. Investigation of the cokernel of $\mathrm{Hom}_R(X,\ )$ is done in homological algebra; it is involved with a functor called $\mathrm{Ext}^1_R(X,\ )$.

There is an analogous result for contravariant Hom functors.

Theorem 7.46. *If*

$$A \xrightarrow{i} B \xrightarrow{p} C \to 0$$

is an exact sequence of R-modules, and if Y is an R-module, then there is an exact sequence

$$0 \to \mathrm{Hom}_R(C, Y) \xrightarrow{p^*} \mathrm{Hom}_R(B, Y) \xrightarrow{i^*} \mathrm{Hom}_R(A, Y).$$

Proof. (i) $\ker p^* = \{0\}$.

If $h \in \ker p^*$, then $h\colon C \to Y$ and $0 = p^*(h) = hp$. Thus, $h(p(b)) = 0$ for all $b \in B$, so that $h(c) = 0$ for all $c \in \mathrm{im}\ p$. Since p is surjective, $\mathrm{im}\ p = C$, and $h = 0$.

(ii) $\mathrm{im}\ p^* \subseteq \ker i^*$.

If $g \in \mathrm{Hom}_R(C, Y)$, then

$$i^* p^*(g) = (pi)^*(g) = 0,$$

because exactness of the original sequence, namely, $\mathrm{im}\, i = \ker p$, implies $pi = 0$.

(iii) $\ker i^* \subseteq \mathrm{im}\ p^*$.

If $g \in \ker i^*$, then $g\colon B \to Y$ and $i^*(g) = gi = 0$. If $c \in C$, then $c = p(b)$ for some $b \in B$, because p is surjective. Define $f\colon C \to Y$ by $f(c) = g(b)$ if $c = p(b)$. Note that f is well-defined: If $p(b) = p(b')$, then $b - b' \in \ker p = \mathrm{im}\, i$, so that $b - b' = i(a)$ for some $a \in A$. Hence,

$$g(b) - g(b') = g(b - b') = gi(a) = 0,$$

because $gi = 0$. The reader may check that f is an R-map. Finally,

$$p^*(f) = fp = g,$$

because if $c = p(b)$, then $g(b) = f(c) = f(p(b))$. Therefore, $g \in \mathrm{im}\ p^*$. •

Example 7.47.
Even if the map $i: A \to B$ in the original exact sequence is assumed to be injective, the functored sequence need not end with "$\to 0$;" that is, $i^*: \operatorname{Hom}_R(B, Y) \to \operatorname{Hom}_R(A, Y)$ may fail be surjective.

We claim that $\operatorname{Hom}_{\mathbb{Z}}(\mathbb{Q}, \mathbb{Z}) = 0$. Suppose that $f: \mathbb{Q} \to \mathbb{Z}$ and $f(a/b) \neq 0$ for some $a/b \in \mathbb{Q}$. If $f(a/b) = m$, then, for all $n > 0$,

$$nf(a/nb) = f(na/nb) = f(a/b) = m.$$

Thus, m is divisible by every positive integer n, and this contradicts the fundamental theorem of arithmetic.

If we apply the functor $\operatorname{Hom}_{\mathbb{Z}}(\ \ , \mathbb{Z})$ to the short exact sequence

$$0 \to \mathbb{Z} \xrightarrow{i} \mathbb{Q} \xrightarrow{p} \mathbb{Q}/\mathbb{Z} \to 0,$$

where i is the inclusion and p is the natural map, then the induced map

$$i^*: \operatorname{Hom}_{\mathbb{Z}}(\mathbb{Q}, \mathbb{Z}) \to \operatorname{Hom}_{\mathbb{Z}}(\mathbb{Z}, \mathbb{Z})$$

cannot be surjective, for $\operatorname{Hom}_{\mathbb{Z}}(\mathbb{Q}, \mathbb{Z}) = \{0\}$ while $\operatorname{Hom}_{\mathbb{Z}}(\mathbb{Z}, \mathbb{Z}) \neq \{0\}$, because it contains $1_{\mathbb{Z}}$. ◄

Definition. A contravariant functor $T: {}_R\mathbf{Mod} \to \mathbf{Ab}$ is called ***left exact*** if exactness of

$$A \xrightarrow{i} B \xrightarrow{p} C \to 0$$

implies exactness of

$$0 \to T(C) \xrightarrow{T(p)} T(B) \xrightarrow{T(i)} T(A).$$

Thus, Theorem 7.46 shows that contravariant Hom functors $\operatorname{Hom}_R(\ \ , Y)$ are left exact functors.[12]

There is a converse of Theorem 7.46; a dual statement holds for covariant Hom functors.

Proposition 7.48. *Let $i: B' \to B$ and $p: B \to B''$ be R-maps, where R is a commutative ring. If, for every R-module M,*

$$0 \to \operatorname{Hom}_R(B'', M) \xrightarrow{p^*} \operatorname{Hom}_R(B, M) \xrightarrow{i^*} \operatorname{Hom}_R(B', M)$$

is an exact sequence, then so is

$$B' \xrightarrow{i} B \xrightarrow{p} B'' \to 0.$$

[12] These functors are called *left exact* because the functored sequence has $0 \to$ on the left.

Proof. (i) p is surjective.

Let $M = B''/\operatorname{im} p$ and let $f: B'' \to B''/\operatorname{im} p$ be the natural map, so that $f \in \operatorname{Hom}(B'', M)$. Then $p^*(f) = fp = 0$, so that $f = 0$, because p^* is injective. Therefore, $B''/\operatorname{im} p = 0$, and p is surjective.

(ii) $\operatorname{im} i \subseteq \ker p$.

Since $i^* p^* = 0$, we have $0 = (pi)^*$. Hence, if $M = B''$ and $g = 1_{B''}$, so that $g \in \operatorname{Hom}(B'', M)$, then $0 = (pi)^* g = gpi = pi$, and so $\operatorname{im} i \subseteq \ker p$.

(iii) $\ker p \subseteq \operatorname{im} i$.

Now choose $M = B/\operatorname{im} i$ and let $h: B \to M$ be the natural map, so that $h \in \operatorname{Hom}(B, M)$. Clearly, $i^* h = hi = 0$, so that exactness of the Hom sequence gives an element $h' \in \operatorname{Hom}_R(B'', M)$ with $p^*(h') = h'p = h$. We have $\operatorname{im} i \subseteq \ker p$, by part (ii); hence, if $\operatorname{im} i \neq \ker p$, there is an element $b \in B$ with $b \notin \operatorname{im} i$ and $b \in \ker p$. Thus, $hb \neq 0$ and $pb = 0$, which gives the contradiction $hb = h'pb = 0$. •

Definition. A covariant functor $T: {}_R\mathbf{Mod} \to \mathbf{Ab}$ is an ***exact functor*** if exactness of

$$0 \to A \xrightarrow{i} B \xrightarrow{p} C \to 0$$

implies exactness of

$$0 \to T(A) \xrightarrow{T(i)} T(B) \xrightarrow{T(p)} T(C) \to 0.$$

An exact contravariant functor is defined similarly.

In the next section, we will see that Hom functors are exact functors for certain choices of modules.

EXERCISES

7.34 If $T: {}_R\mathbf{Mod} \to \mathbf{Ab}$ is an additive functor, of either variance, prove that $T(0) = 0$, where 0 denotes either a zero module or a zero morphism.

7.35 Give an example of a covariant functor that does not preserve coproducts.
Hint. Use Exercise 7.21(iii) on page 458.

7.36 Let $\mathcal{A} \xrightarrow{S} \mathcal{B} \xrightarrow{T} \mathcal{C}$ be functors. Prove that the composite $\mathcal{A} \xrightarrow{TS} \mathcal{C}$ is a functor that is covariant if the variances of S and T are the same, and contravariant if the variances of S and T are different.

7.37 (i) Prove that there is a functor on **CommRings** defined on objects by $R \mapsto R[x]$, and on morphisms $f: R \to S$ by $r \mapsto f(r)$ (that is, in the formal notation for elements of $R[x]$, $(r, 0, 0, \cdots) \mapsto (f(r), 0, 0, \cdots)$.

(ii) Prove that there is a functor on **Dom**, the category of all domains, defined on objects by $R \mapsto \operatorname{Frac}(R)$, and on morphisms $f: R \to S$ by $r/1 \mapsto f(r)/1$.

7.38 Prove that there is a functor **Groups** $\to$ **Ab** taking each group G to G/G', where G' is its commutator subgroup.

7.39 (i) If X is a set and k is a field, define the vector space k^X to be the set of all functions $X \to k$ under pointwise operations. Prove that there is a functor $F\colon$ **Sets** $\to {}_k$**Mod** with $F(X) = k^X$.

(ii) If X is a set, define $F(X)$ to be the free group with basis X. Prove that there is a functor $F\colon$ **Sets** $\to$ **Groups** with $F\colon X \mapsto F(X)$.

7.4 FREE MODULES, PROJECTIVES, AND INJECTIVES

The simplest modules are free modules and, as for groups, every module is a quotient of a free module; that is, every module has a presentation by generators and relations. Projective modules are generalizations of free modules, and they, too, turn out to be useful. We define injective modules, as duals of projectives, but their value cannot be appreciated until Chapter 10, when we discuss homological algebra. In the meantime, we will see here that injective $\mathbb{Z}$-modules are quite familiar.

Definition. An R-module F is called a ***free R-module*** if F is isomorphic to a direct sum of copies of R: that is, there is a (possibly infinite) index set I with

$$F = \sum_{i \in I} R_i,$$

where $R_i = \langle b_i \rangle \cong R$ for all i. We call $B = \{b_i : i \in I\}$ a ***basis*** of F.

A free $\mathbb{Z}$-module is a free abelian group, and every commutative ring R, when considered as a module over itself, is itself a free R-module.

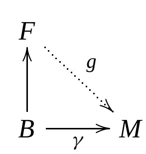
From our discussion of direct sums, we know that each $m \in F$ has a unique expression of the form

$$m = \sum_{i \in I} r_i b_i,$$

where $r_i \in R$ and almost all $r_i = 0$. A basis of a free module has a strong resemblence to a basis of a vector space. Indeed, it is easy to see that a vector space V over a field k is a free k-module, and that the two notions of basis coincide in this case.

There is a straightforward generalization of Theorem 3.92 from finite-dimensional vector spaces to arbitrary free modules (in particular, to infinite-dimensional vector spaces).

Proposition 7.49. *Let F be a free R-module, and let $B = \{b_i : i \in I\}$ be a basis of F. If M is any R-module and if $\gamma\colon B \to M$ is any function, then there exists a unique R-map $g\colon F \to M$ with $g(b_i) = \gamma(b_i)$ for all $i \in I$.*



Proof. Every element $v \in F$ has a unique expression of the form

$$v = \sum_{i \in I} r_i b_i,$$

where $r_i \in R$ and almost all $r_i = 0$. Define $g\colon F \to M$ by

$$g(v) = \sum_{i \in I} r_i \gamma(b_i). \quad \bullet$$

Here is a fancy proof of this result. By Proposition 7.30, a free module F is the coproduct of $\{\langle b_i \rangle : i \in I\}$, with injections α_i mapping $r_i b_i$ to the vector having $r_i b_i$ in the ith coordinate and 0's elsewhere. As for any coproduct, there is a unique map $\theta\colon F \to M$ with $\theta(b_i) = \gamma(b_i)$. The maps θ and g agree on each element of the basis B, so that $\theta = g$.

Definition. The number of elements in a basis is called the ***rank*** of F.

Of course, rank is the analog of dimension. The next proposition shows that rank is well-defined.

Proposition 7.50.

(i) *If R is a nonzero commutative ring, then any two bases of a free R-module F have the same cardinality; that is, the same number of elements.*

(ii) *If R is a nonzero commutative ring, then free R-modules F and F' are isomorphic if and only if* $\operatorname{rank}(F) = \operatorname{rank}(F')$.

Proof. (i) Choose a maximal ideal I in R (which exists, by Theorem 6.46). If X is a basis of the free R-module F, then Exercise 7.6 on page 440 shows that the set of cosets $\{v + IF : v \in X\}$ is a basis of the vector space F/IF over the field R/I. If Y is another basis of F, then the same argument gives $\{u + IF : u \in Y\}$ a basis of F/IF. But any two bases of a vector space have the same size (which is the dimension of the space), and so $|X| = |Y|$, by Theorem 6.51.

(ii) Let X be a basis of F, let X' be a basis of F', and let $\gamma\colon X \to X'$ be a bijection. Composing γ with the inclusion $X' \to F'$, we may assume that $\gamma\colon X \to F'$. By Proposition 7.49, there is a unique R-map $\varphi\colon F \to F'$ extending γ. Similarly, we may regard $\gamma^{-1}\colon X' \to X$ as a function $X' \to F$, and there is a unique $\psi\colon F' \to F$ extending γ^{-1}. Finally, both $\psi\varphi$ and 1_F extend 1_X, so that $\psi\varphi = 1_F$. Similarly, the other composite is $1_{F'}$, and so $\varphi\colon F \to F'$ is an isomorphism. (The astute reader will notice a strong resemblance of this proof to the uniqueness of a solution to a universal mapping problem.)

Conversely, suppose that $\varphi\colon F \to F'$ is an isomorphism. If $\{v_i : i \in I\}$ is a basis of F, then it is easy to see that $\{\varphi(v_i) : i \in I\}$ is a basis of F'. But any two bases of the free module F' have the same size, namely, $\operatorname{rank}(F')$, by part (i). Hence, $\operatorname{rank}(F') = \operatorname{rank}(F)$. $\bullet$

The next proposition will enable us to use free modules to describe arbitrary modules.

Proposition 7.51. *Every R-module M is a quotient of a free R-module F. Moreover, M is finitely generated if and only if F can be chosen to be finitely generated.*

Proof. Let F be the direct sum of $|M|$ copies of R (so F is a free module), and let $\{x_m : m \in M\}$ be a basis of F. By Proposition 7.49, there is an R-map $g \colon F \to M$ with $g(x_m) = m$ for all $m \in M$. Obviously, g is a surjection, and so $F/\ker g \cong M$.

If M is finitely generated, then $M = \langle m_1, \ldots, m_n \rangle$. If we choose F to be the free R-module with basis $\{x_1, \ldots, x_n\}$, then the map $g \colon F \to M$ with $g(x_i) = m_i$ is a surjection, for

$$\operatorname{im} g = \langle g(x_1), \ldots, g(x_n) \rangle = \langle m_1, \ldots, m_n \rangle = M.$$

The converse is obvious, for any image of a finitely generated module is itself finitely generated •

The last proposition can be used to construct modules with prescribed properties. For example, let us consider $\mathbb{Z}$-modules (i.e., abelian groups). The group $\mathbb{Q}/\mathbb{Z}$ contains an element a of order 2 satisfying the equations $a = 2^n a_n$ for all $n \geq 1$; take $a = \frac{1}{2} + \mathbb{Z}$ and $a_n = 1/2^{n+1} + \mathbb{Z}$. Of course, $\operatorname{Hom}_{\mathbb{Z}}(\mathbb{Q}, \mathbb{Q}/\mathbb{Z}) \neq \{0\}$ because it contains the natural map. Is there an abelian group G with $\operatorname{Hom}_{\mathbb{Z}}(\mathbb{Q}, G) = \{0\}$ that contains an element a of order 2 satisfying the equations $a = 2^n a_n$ for all $n \geq 1$? Let F be the free abelian group with basis

$$\{a, b_1, b_2, \ldots, b_n, \ldots\}$$

and relations

$$\{2a, a - 2^n b_n, n \geq 1\};$$

that is, let K be the subgroup of F generated by $\{2a, a - 2^n b_n, n \geq 1\}$. Exercise 7.48 on page 487 asks the reader to verify that $G = F/K$ satisfies the desired properties. This construction is a special case of defining an R-module by *generators and relations* (as we have already done for groups).

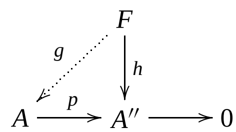
Definition. Let $\mathcal{X} = \{x_i : i \in I\}$ be a basis of a free R-module F, and let $\mathcal{R} = \{\sum_i r_{ji} x_i : j \in J\}$ be a subset of F. If K is the submodule of F generated by $\mathcal{R}$, then we say that the module $M = F/K$ has ***generators*** $\mathcal{X}$ and ***relations*** $\mathcal{R}$.[13] We also say that the ordered pair $(\mathcal{X}|\mathcal{R})$ is a ***presentation*** of M.

We will return to presentations at this end of the section, but let us now focus on the key property of bases, Lemma 7.49 (which holds for free modules as well as for vector spaces), in order to get a theorem about free modules that does not mention bases.

Theorem 7.52. *If R is a commutative ring and F is a free R-module, then for every surjection $p \colon A \to A''$ and each $h \colon F \to A''$, there exists a homomorphism g making the*

[13] A module is called *free* because it has no entangling relations.

following diagram commute:



Proof. Let $\{b_i : i \in I\}$ be a basis of F. Since p is surjective, there is $a_i \in A$ with $p(a_i) = h(b_i)$ for all i. By Proposition 7.49, there is an R-map $g\colon F \to A$ with

$$g(b_i) = a_i \text{ for all } i.$$

Now $pg(b_i) = p(a_i) = h(b_i)$, so that pg agrees with h on the basis $\{b_i : i \in I\}$; it follows that $pg = h$ on $\langle\{b_i : i \in I\}\rangle = F$; that is, $pg = h$. •

Definition. We call a map $g\colon F \to A$ with $pg = h$ (in the diagram in Theorem 7.52) a ***lifting*** of h.

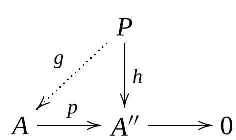
If C is any, not necessarily free, module, then a lifting g of h, should one exist, need not be unique. Since $pi = 0$, where $i\colon \ker p \to A$ is the inclusion, other liftings are $g + if$ for any $f \in \mathrm{Hom}_R(C, \ker p)$. Indeed, this is obvious from the exact sequence

$$0 \to \mathrm{Hom}(C, \ker p) \xrightarrow{i_*} \mathrm{Hom}(C, A) \xrightarrow{p_*} \mathrm{Hom}(C, A'').$$

Any two liftings of h differ by a map in $\ker p_* = \mathrm{im}\, i_* \subseteq \mathrm{Hom}(C, A)$.

We now promote this (basis-free) property of free modules to a definition.

Definition. A module P is ***projective*** if, whenever p is surjective and h is any map, there exists a lifting g; that is, there exists a map g making the following diagram commute:



We know that every free module is projective; is every projective R-module free? We shall see that the answer to this question depends on the ring R. Note that if projective R-modules happen to be free, then free modules are characterized without having to refer to a basis.

Let us now see that projective modules arise in a natural way. We know that the Hom functors are left exact; that is, for any module P, applying $\mathrm{Hom}_R(P, \ )$ to an exact sequence

$$0 \to A' \xrightarrow{i} A \xrightarrow{p} A''$$

gives an exact sequence

$$0 \to \mathrm{Hom}_R(P, A') \xrightarrow{i_*} \mathrm{Hom}_R(P, A) \xrightarrow{p_*} \mathrm{Hom}_R(P, A'').$$

Proposition 7.53. *A module P is projective if and only if* $\mathrm{Hom}_R(P,\ )$ *is an exact functor.*

Remark. Since $\mathrm{Hom}_R(P,\ )$ is a left exact functor, the thrust of the proposition is that p_* is surjective whenever p is surjective. ◂

Proof. If P is projective, then given $h\colon P \to A''$, there exists a lifting $g\colon P \to A$ with $pg = h$. Thus, if $h \in \mathrm{Hom}_R(P, A'')$, then $h = pg = p_*(g) \in \mathrm{im}\, p_*$, and so p_* is surjective. Hence, $\mathrm{Hom}(P,\ )$ is an exact functor.

For the converse, assume that $\mathrm{Hom}(P,\ )$ is an exact functor, so that p_* is surjective: If $h \in \mathrm{Hom}_R(P, A'')$, there exists $g \in \mathrm{Hom}_R(P, A)$ with $h = p_*(g) = pg$. This says that given p and h, there exists a lifting g making the diagram commute; that is, P is projective. •

Proposition 7.54. *A module P is projective if and only if every short exact sequence*

$$0 \to A \xrightarrow{i} B \xrightarrow{p} P \to 0$$

is split.

Proof. If P is projective, then there exists $j\colon P \to B$ making the following diagram commute; that is, $pj = 1_P$.

$$\begin{array}{ccccc} & & P & & \\ & {\scriptstyle j}\swarrow & \downarrow{\scriptstyle 1_P} & & \\ B & \xrightarrow{p} & P & \longrightarrow & 0 \end{array}$$

Corollary 7.17 now gives the result.

Conversely, assume that every short exact sequence ending with P splits. Consider the diagram

$$\begin{array}{ccccc} & & P & & \\ & & \downarrow{\scriptstyle f} & & \\ B & \xrightarrow{p} & C & \longrightarrow & 0 \end{array}$$

with p surjective. Now form the pullback

$$\begin{array}{ccccc} D & \underset{j}{\overset{\alpha}{\rightleftarrows}} & P & & \\ {\scriptstyle \beta}\downarrow & & \downarrow{\scriptstyle f} & & \\ B & \xrightarrow{p} & C & \longrightarrow & 0 \end{array}$$

By Exercise 7.28 on page 459, surjectivity of p in the pullback diagram gives surjectivity of α. By hypothesis, there is a map $j\colon P \to D$ with $\alpha j = 1_P$. Define $g\colon P \to B$ by $g = \beta j$. We check:

$$pg = p\beta j = f\alpha j = f1_P = f.$$

Therefore, P is projective. •

We restate one half of this proposition so that the word *exact* is not mentioned.

Corollary 7.55. *Let A be a submodule of a module B. If B/A is projective, then there is a submodule C of B with $C \cong B/A$ and $B = A \oplus C$.*

Theorem 7.56. *An R-module P is projective if and only if P is a direct summand of a free R-module.*

Proof. Assume that P is projective. By Proposition 7.51, every module is a quotient of a free module. Thus, there is a free module F and a surjection $g\colon F \to P$, and so there is an exact sequence

$$0 \to \ker g \to F \xrightarrow{g} P \to 0.$$

Proposition 7.54 now shows that P is a direct summand of F.

Suppose that P is a direct summand of a free module F, so there are maps $q\colon F \to P$ and $j\colon P \to F$ with $qj = 1_P$. Now consider the diagram

$$\begin{array}{ccccc} F & \underset{j}{\overset{q}{\rightleftarrows}} & P & & \\ {\scriptstyle h}\downarrow & & \downarrow{\scriptstyle f} & & \\ B & \xrightarrow{p} & C & \longrightarrow & 0, \end{array}$$

where p is surjective. The composite fq is a map $F \to C$; since F is free, it is projective, and so there is a map $h\colon F \to B$ with $ph = fq$. Define $g\colon P \to B$ by $g = hj$. It remains to prove that $pg = f$. But

$$pg = phj = fqj = f1_P = f. \quad \bullet$$

Actually, the second half of the proof shows that any direct summand of a projective module is itself projective.

We can now give an example of a commutative ring R and a projective R-module that is not free.

Example 7.57.
The ring $R = \mathbb{I}_6$ is the direct sum of two ideals:

$$\mathbb{I}_6 = J \oplus I,$$

where

$$J = \{[0], [2], [4]\} \cong \mathbb{I}_3 \text{ and } I = \{[0], [3]\} \cong \mathbb{I}_2.$$

Now $\mathbb{I}_6$ is a free module over itself, and so J and I, being direct summands of a free module, are projective $\mathbb{I}_6$-modules. Neither J nor I can be free, however. After all, a (finitely generated) free $\mathbb{I}_6$-module F is a direct sum of, say, n copies of $\mathbb{I}_6$, and so F has 6^n elements. Therefore, J is too small to be free, for it has only three elements. ◂

Describing projective R-modules is a problem very much dependent on the ring R. In Chapter 9, for example, we will prove that if R is a PID, then every submodule of a free module is itself free. It will then follow from Theorem 7.56 that every projective R-module is free in this case. A much harder result is that if $R = k[x_1, \ldots, x_n]$ is the polynomial ring in n variables over a field k, then every projective R-module is also free; this theorem, implicitly conjectured[14] by J.-P. Serre, was proved, independently, by D. Quillen and by A. Suslin (see Rotman, *An Introduction to Homological Algebra*, pages 138–145, for a proof). There is a proof of the Quillen-Suslin theorem using Gröbner bases, due to N. Fitchas, A. Galligo, and B. Sturmfels.

There are domains having projective modules that are not free. For example, if R is the ring of all the algebraic integers in an *algebraic number field* (that is, an extension of $\mathbb{Q}$ of finite degree), then every ideal in R is a projective R-module. There are such rings R that are not PIDs, and any ideal in R that is not principal is a projective module that is not free (we will see this in Chapter 11 when we discuss Dedekind rings).

Here is another characterization of projective modules. Note that if A is a free R-module with basis $\{a_i : i \in I\} \subseteq A$, then each $x \in A$ has a unique expression $x = \sum_{i \in I} r_i a_i$, and so there are R-maps $\varphi_i : A \to R$ given by $\varphi_i : x \mapsto r_i$.

Proposition 7.58. *An R-module A is projective if and only if there exist elements $\{a_i : i \in I\} \subseteq A$ and R-maps $\{\varphi_i : A \to R : i \in I\}$ such that*

(i) *for each $x \in A$, almost all $\varphi_i(x) = 0$;*

(ii) *for each $x \in A$, we have $x = \sum_{i \in I} (\varphi_i x) a_i$.*

Moreover, A is generated by $\{a_i : i \in I\} \subseteq A$ in this case.

Proof. If A is projective, there is a free R-module F and a surjective R-map $\psi : F \to A$. Since A is projective, there is an R-map $\varphi : A \to F$ with $\psi\varphi = 1_A$, by Proposition 7.54. Let $\{e_i : i \in I\}$ be a basis of F, and define $a_i = \psi(e_i)$. Now if $x \in A$, then there is a unique expression $\varphi(x) = \sum_i r_i e_i$, where $r_i \in R$ and almost all $r_i = 0$. Define $\varphi_i : A \to R$ by $\varphi_i(x) = r_i$. Of course, given x, we have $\varphi_i(x) = 0$ for almost all i. Since ψ is surjective, A is generated by $\{a_i = \psi(e_i) : i \in I\}$. Finally,

$$\begin{aligned} x = \psi\varphi(x) &= \psi\left(\sum r_i e_i\right) \\ &= \sum r_i \psi(e_i) = \sum (\varphi_i x)\psi(e_i) = \sum (\varphi_i x) a_i. \end{aligned}$$

Conversely, given $\{a_i : i \in I\} \subseteq A$ and a family of R-maps $\{\varphi_i : A \to R : i \in I\}$ as in the statement, define F to be the free R-module with basis $\{e_i : i \in I\}$, and define an R-map $\psi : F \to A$ by $\psi : e_i \mapsto a_i$. It suffices to find an R-map $\varphi : A \to F$ with $\psi\varphi = 1_A$, for then A is (isomorphic to) a retract (i.e., A is a direct summand of F), and hence A is projective. Define φ by $\varphi(x) = \sum_i (\varphi_i x) e_i$, for $x \in A$. The sum is finite, by

[14] On page 243 of “Faisceaux Algèbriques Cohérents,” *Annals of Mathematics* 61 (1955), 197–278, Serre writes “... on ignore s’il existe des A-modules projectifs de type fini qui ne soient pas libres.” Here, $A = k[x_1, \ldots, x_n]$.

condition (i), and so φ is well-defined. By condition (ii),

$$\psi\varphi(x) = \psi \sum(\varphi_i x)e_i = \sum(\varphi_i x)\psi(e_i) = \sum(\varphi_i x)a_i = x;$$

that is, $\psi\varphi = 1_A$. •

Definition. If A is an R-module, then a subset $\{a_i : i \in I\} \subseteq A$ and a family of R-maps $\{\varphi_i \colon A \to R : i \in I\}$ satisfying the condition in Proposition 7.58 is called a ***projective basis***.

An interesting application of projective bases is due to R. Bkouche. Let X be a locally compact Hausdorff space, let $C(X)$ be the ring of all continuous real-valued functions on X, and let J be the ideal in $C(X)$ consisting of all such functions having compact support. Then X is a paracompact space if and only if J is a projective $C(X)$-module.

Remark. The definition of projective module can be used to define a projective object in any category (we do not assert that such objects always exist), if we can translate *surjection* into the language of categories. One candidate arises from Exercise 7.18 on page 441, but we shall see now that defining surjections in arbitrary categories is not so straightforward.

Definition. A morphism $\varphi \colon B \to C$ in a category $\mathcal{C}$ is an ***epimorphism*** if φ can be canceled from the right; that is, for all objects D and all morphisms $h \colon C \to D$ and $k \colon C \to D$, we have $h\varphi = k\varphi$ implies $h = k$.

$$B \xrightarrow{\varphi} C \underset{k}{\overset{h}{\rightrightarrows}} D$$

Now Exercise 7.18 on page 441 shows that epimorphisms in ${}_R\mathbf{Mod}$ are precisely the surjections, and Exercises 7.45 on page 487 and 7.19 on page 441 show that epimorphisms in **Sets** and in **Groups**, respectively, are also surjections. However, in **CommRings**, it is easy to see that if R is a domain, then the ring homomorphism $\varphi \colon R \to \mathrm{Frac}(R)$, given by $r \mapsto r/1$, is an epimorphism; if A is a commutative ring and $h, k \colon \mathrm{Frac}(R) \to A$ are ring homomorphisms that agree on R, then $h = k$. But φ is not a surjective function if R is not a field. A similar phenomenon occurs in **Top**. If $f \colon X \to Y$ is a continuous map with $\mathrm{im}\, f$ a dense subspace of Y, then f is an epimorphism, because any two continuous functions agreeing on a dense subspace must be equal.

There is a similar problem with ***monomorphisms***, a generalization of injections to arbitrary categories: A category whose objects have underlying sets may have monomorphisms whose underlying function is not an injection. ◀

Let us return to presentations of modules.

Definition. An R-module M is ***finitely presented*** if it has a presentation $(\mathcal{X}|\mathcal{R})$ in which both $\mathcal{X}$ and $\mathcal{R}$ are finite.

If M is finitely presented, there is a short exact sequence

$$0 \to K \to F \to M \to 0,$$

where F is free and both K and F are finitely generated. Equivalently, M is finitely presented if there is an exact sequence

$$F' \to F \to M \to 0,$$

where both F' and F are finitely generated free modules (just map a finitely generated free module F' onto K). Note that the second exact sequence does not begin with "$0 \to$."

Proposition 7.59. *If R is a commutative noetherian ring, then every finitely generated R-module is finitely presented.*

Proof. If M is a finitely generated R-module, then there is a finitely generated free R-module F and a surjection $\varphi\colon F \to M$. Since R is noetherian, Proposition 7.23 says that every submodule of F is finitely generated. In particular, $\ker\varphi$ is finitely generated, and so M is finitely presented. •

Every finitely presented module is finitely generated, but we will soon see that the converse may be false. We begin by comparing two presentations of a module (we generalize a bit by replacing free modules by projectives).

Proposition 7.60 (Schanuel's Lemma). *Given exact sequences*

$$0 \to K \xrightarrow{i} P \xrightarrow{\pi} M \to 0$$

and

$$0 \to K' \xrightarrow{i'} P' \xrightarrow{\pi'} M \to 0,$$

where P and P' are projective, then there is an isomorphism

$$K \oplus P' \cong K' \oplus P.$$

Proof. Consider the diagram with exact rows

$$\begin{array}{ccccccccc}
0 & \longrightarrow & K & \xrightarrow{i} & P & \xrightarrow{\pi} & M & \longrightarrow & 0 \\
 & & \downarrow{\scriptstyle \alpha} & & \downarrow{\scriptstyle \beta} & & \downarrow{\scriptstyle 1_M} & & \\
0 & \longrightarrow & K' & \xrightarrow{i'} & P' & \xrightarrow{\pi'} & M & \longrightarrow & 0
\end{array}$$

Since P is projective, there is a map $\beta\colon P \to P'$ with $\pi'\beta = \pi$; that is, the right square in the diagram commutes. We now show that there is a map $\alpha\colon K \to K'$ making the other square commute. If $x \in K$, then $\pi'\beta i x = \pi i x = 0$, because $\pi i = 0$. Hence, $\beta i x \in \ker\pi' = \operatorname{im} i'$; thus, there is $x' \in K'$ with $i'x' = \beta i x$; moreover, x' is unique

because i' is injective. Therefore, $\alpha\colon x \mapsto x'$ is a well-defined function $\alpha\colon K \to K'$ that makes the first square commute. The reader can show that α is an R-map.

This commutative diagram with exact rows gives an exact sequence

$$0 \to K \xrightarrow{\theta} P \oplus K' \xrightarrow{\psi} P' \to 0,$$

where $\theta\colon x \mapsto (ix, \alpha x)$ and $\psi\colon (u, x') \mapsto \beta u - i'x'$, for $x \in K$, $u \in P$, and $x' \in K'$. Exactness of this sequence is a straightforward calculation that is left to the reader; this sequence splits because P' is projective. $\bullet$

Corollary 7.61. *If M is finitely presented and*

$$0 \to K \to F \to M \to 0$$

is an exact sequence, where F is a finitely generated free module, then K is finitely generated.

Proof. Since M is finitely presented, there is an exact sequence

$$0 \to K' \to F' \to M \to 0$$

with F' free and with both F' and K' finitely generated. By Schanuel's lemma, $K \oplus F' \cong K' \oplus F$. Now $K' \oplus F$ is finitely generated because both summands are, so that the left side is also finitely generated. But K, being a summand, is also a homomorphic image of $K \oplus F'$, and hence it is finitely generated. $\bullet$

We can now give an example of a finitely generated module that is not finitely presented.

Example 7.62.

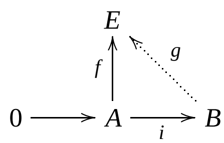
Let R be a commutative ring that is not noetherian; that is, R contains an ideal I that is not finitely generated (see Example 6.39). We claim that the R-module $M = R/I$ is finitely generated but not finitely presented. Of course, M is finitely generated; it is even cyclic. If M were finitely presented, then there would be an exact sequence $0 \to K \to F \to M \to 0$ with F free and both K and F finitely generated. Comparing this with the exact sequence $0 \to I \to R \to M \to 0$, as in Corollary 7.61, gives I finitely generated, a contradiction. Therefore, M is not finitely presented. ◂

There is another type of module that also turns out to be interesting.

Definition. If E is a module for which the contravariant Hom functor $\mathrm{Hom}_R(\ \ , E)$ is an exact functor—that is, if $\mathrm{Hom}_R(\ \ , E)$ preserves all short exact sequences—then E is called an ***injective*** module.

The next proposition is the dual of Proposition 7.53.

Proposition 7.63. *A module E is injective if and only if a dotted arrow always exists making the following diagram commute whenever i is an injection*:



In words, every homomorphism from a submodule into E can always be extended to a homomorphism from the big module into E.

Remark. Since $\mathrm{Hom}_R(\ \ , E)$ is a left exact contravariant functor, the thrust of the proposition is that i^* is surjective whenever i is injective.

Injective modules are duals of projective modules in that both of these terms are characterized by diagrams, and the diagram for injectivity is the diagram for projectivity having all arrows reversed. ◂

Proof. If E is injective, then $\mathrm{Hom}(\ , E)$ is an exact functor, so that i^* is surjective. Therefore, if $f \in \mathrm{Hom}_R(A, E)$, there exists $g \in \mathrm{Hom}_R(B, E)$ with $f = i^*(g) = gi$; that is, the diagram commutes.

For the converse, if E satisfies the diagram condition, then given $f : A \to E$, there exists $g : B \to E$ with $gi = f$. Thus, if $f \in \mathrm{Hom}_R(A, E)$, then $f = gi = i^*(g) \in \mathrm{im}\, i^*$, and so i^* is surjective. Hence, $\mathrm{Hom}(\ , E)$ is an exact functor, and so E is injective. •

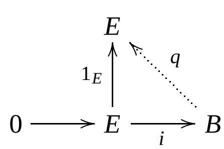
The next result is the dual of Proposition 7.54.

Proposition 7.64. *A module E is injective if and only if every short exact sequence*

$$0 \to E \xrightarrow{i} B \xrightarrow{p} C \to 0$$

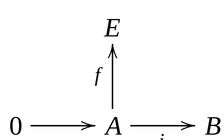
is split.

Proof. If E is injective, then there exists $q : B \to E$ making the following diagram commute; that is, $qi = 1_E$.



Exercise 7.17 on page 441 now gives the result.

Conversely, assume every exact sequence beginning with E splits. The pushout of



is the diagram

$$\begin{array}{ccccc} & & E & \xrightarrow{\alpha} & D \\ & & {\scriptstyle f}\uparrow & & \uparrow{\scriptstyle \beta} \\ 0 & \longrightarrow & A & \xrightarrow[i]{} & B \end{array}$$

By Exercise 7.28 on page 459, the map α is an injection, so that

$$0 \to E \to D \to \operatorname{coker}\alpha \to 0$$

splits; that is, there is $q\colon D \to E$ with $q\alpha = 1_E$. If we define $g\colon B \to E$ by $g = q\beta$, then the original diagram commutes:

$$gi = q\beta i = q\alpha f = 1_E f = f.$$

Therefore, E is injective. •

This proposition can be restated without mentioning the word *exact*.

Corollary 7.65. *If an injective module E is a submodule of a module M, then E is a direct summand of M: There is a submodule S of M with $S \cong M/E$ and $M = E \oplus S$.*

Proposition 7.66. *If $\{E_i : i \in I\}$ is a family of injective modules, then $\prod_{i \in I} E_i$ is also an injective module.*

Proof. Consider the diagram

$$\begin{array}{ccccc} & & E & & \\ & & {\scriptstyle f}\uparrow & & \\ 0 & \longrightarrow & A & \xrightarrow[\kappa]{} & B, \end{array}$$

where $E = \prod E_i$. Let $p_i\colon E \to E_i$ be the ith projection. Since E_i is injective, there is $g_i\colon B \to E_i$ with $g_i\kappa = p_i f$. Now define $g\colon B \to E$ by $g\colon b \mapsto (g_i(b))$. The map g does extend f, for if $b = \kappa a$, then

$$g(\kappa a) = (g_i(\kappa a)) = (p_i f a) = fa,$$

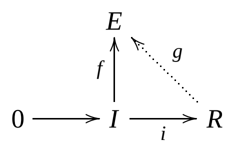
because $x = (p_i x)$ is true for every x in the product. •

Corollary 7.67. *A finite direct sum of injective modules is injective.*

Proof. The direct sum of finitely many modules coincides with the direct product. •

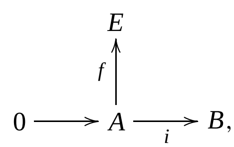
A useful result is the following theorem due to R. Baer.

Theorem 7.68 (Baer Criterion). *An R-module E is injective if and only if every R-map $f: I \to E$, where I is an ideal in R, can be extended to R.*



Proof. Since any ideal I is a submodule of R, the existence of an extension g of f is just a special case of the definition of injectivity of E.

Suppose we have the diagram



where A is a submodule of a module B. For notational convenience, let us assume that i is the inclusion [this assumption amounts to permitting us to write a instead of $i(a)$ whenever $a \in A$]. We are going to use Zorn's lemma on approximations to an extension of f. More precisely, let X be the set of all ordered pairs (A', g'), where $A \subseteq A' \subseteq B$ and $g': A' \to E$ extends f; that is, $g'|A = f$. Note that $X \neq \varnothing$ because $(A, f) \in X$. Partially order X by defining

$$(A', g') \preceq (A'', g'')$$

to mean $A' \subseteq A''$ and g'' extends g'. The reader may supply the argument that Zorn's lemma applies, and so there exists a maximal element (A_0, g_0) in X. If $A_0 = B$, we are done, and so we may assume that there is some $b \in B$ with $b \notin A_0$.

Define

$$I = \{r \in R: rb \in A_0\}.$$

It is easy to see that I is an ideal in R. Define $h: I \to E$ by

$$h(r) = g_0(rb).$$

By hypothesis, there is a map $h^*: R \to E$ extending h. Finally, define $A_1 = A_0 + \langle b \rangle$ and $g_1: A_1 \to E$ by

$$g_1(a_0 + rb) = g_0(a_0) + rh^*(1),$$

where $a_0 \in A_0$ and $r \in R$.

Let us show that g_1 is well-defined. If $a_0 + rb = a_0' + r'b$, then $(r - r')b = a_0' - a_0 \in A_0$; it follows that $r - r' \in I$. Therefore, $g_0((r - r')b)$ and $h(r - r')$ are defined, and we have

$$g_0(a_0' - a_0) = g_0((r - r')b) = h(r - r') = h^*(r - r') = (r - r')h^*(1).$$

Thus, $g_0(a_0') - g_0(a_0) = rh^*(1) - r'h^*(1)$ and $g_0(a_0') + r'h^*(1) = g_0(a_0) + rh^*(1)$, as desired. Clearly, $g_1(a_0) = g_0(a_0)$ for all $a_0 \in A_0$, so that the map g_1 extends g_0. We conclude that $(A_0, g_0) \prec (A_1, g_1)$, contradicting the maximality of (A_0, g_0). Therefore, $A_0 = B$, the map g_0 is a lifting of f, and E is injective. •

Are arbitrary direct sums of injective modules injective?

Proposition 7.69. *If R is noetherian and $\{E_i : i \in I\}$ is a family of injective R-modules, then $\sum_{i \in I} E_i$ is an injective module.*

Proof. By the Baer criterion, Theorem 7.68, it suffices to complete the diagram

$$\begin{array}{ccccc} & & \sum_{i\in I} E_i & & \\ & & \uparrow{\scriptstyle f} & & \\ 0 & \longrightarrow & J & \xrightarrow[\kappa]{} & R, \end{array}$$

where J is an ideal in R. Since R is noetherian, J is finitely generated, say, $J = (a_1, \ldots, a_n)$. For $k = 1, \ldots, n$, $f(a_k) \in \sum_{i \in I} E_i$ has only finitely many nonzero coordinates, occurring, say, at indices in $S(a_k) \subseteq I$. Thus, $S = \bigcup_{k=1}^{n} S(a_k)$ is a finite set, and so $\operatorname{im} f \subseteq \sum_{i \in S} E_i$; by Corollary 7.67, this finite sum is injective. Hence, there is an R-map $g' \colon R \to \sum_{i \in S} E_i$ extending f. Composing g' with the inclusion of $\sum_{i \in S} E_i$ into $\sum_{i \in I} E_i$ completes the given diagram. •

It is a theorem of H. Bass that the converse of Proposition 7.69 is true: If every direct sum of injective R-modules is injective, then R is noetherian (see Theorem 8.105).

We can now give some examples of injective modules.

Proposition 7.70. *If R is a domain, then $Q = \mathrm{Frac}(R)$ is an injective R-module.*

Proof. By Baer's criterion, it suffices to extend an R-map $f : I \to Q$, where I is an ideal in R, to all of R. Note first that if $a, b \in I$ are nonzero, then $af(b) = f(ab) = bf(a)$, so that

$$f(a)/a = f(b)/b \text{ in } Q \text{ for all nonzero } a, b \in I;$$

let $c \in Q$ denote their common value (note how I being an ideal is needed to define c: the product ab must be defined, and either factor can be taken outside the parentheses). Define $g \colon R \to Q$ by

$$g(r) = rc$$

for all $r \in R$. It is obvious that g is an R-map. To see that g extends f, suppose that $a \in I$; then

$$g(a) = ac = af(a)/a = f(a).$$

It now follows from Baer's criterion that Q is an injective R-module. •

Definition. If R is a domain, then an R-module D is ***divisible*** if, for each $d \in D$ and every nonzero $r \in R$, there exists $d' \in D$ with $d = rd'$.

Example 7.71.
Let R be a domain.
(i) Frac(R) is a divisible R-module.

(ii) Every direct sum of divisible R-modules is divisible. Hence, every vector space over Frac(R) is a divisible R-module.

(iii) Every quotient of a divisible R-module is divisible. ◄

Lemma 7.72. *If R is a domain, then every injective R-module E is divisible.*

Proof. Assume that E is injective. Let $e \in E$ and let $r_0 \in R$ be nonzero; we must find $x \in E$ with $e = r_0 x$. Define $f : (r_0) \to E$ by $f(rr_0) = re$ (note that f is well-defined: Since R is a domain, $rr_0 = r'r_0$ implies $r = r'$). Since E is injective, there exists $h : R \to E$ extending f. In particular,

$$e = f(r_0) = h(r_0) = r_0 h(1),$$

so that $x = h(1)$ is the element in E required by the definition of divisible. •

We now prove that the converse of Lemma 7.72 is true for PIDs. Proposition 11.111 shows that a domain R is a Dedekind ring (defined in the last chapter) if and only if every divisible R-module is injective.

Corollary 7.73. *If R is a* PID*, then an R-module E is injective if and only if it is divisible.*

Proof. Assume that E is divisible. By the Baer criterion, Theorem 7.68, it suffices to extend maps $f : I \to E$ to all of R. Since R is a PID, I is principal; say, $I = (r_0)$ for some $r_0 \in I$. Since E is divisible, there exists $e \in E$ with $r_0 e = f(r_0)$. Define $h : R \to E$ by $h(r) = re$. It is easy to see that h is an R-map extending f, and so E is injective. •

Remark. There are domains for which divisible modules are not injective; indeed, there are domains for which a quotient of an injective module need not be injective. ◄

Example 7.74.
In light of Example 7.71, the following abelian groups are injective $\mathbb{Z}$-modules:

$$\mathbb{Q}, \quad \mathbb{R}, \quad \mathbb{C}, \quad \mathbb{Q}/\mathbb{Z}, \quad \mathbb{R}/\mathbb{Z}, \quad S^1,$$

where S^1 is the circle group; that is, the multiplicative group of all complex numbers z with $|z| = 1$. ◄

Proposition 7.51 says that, over any ring, every module is a quotient of a projective module (actually, it is a stronger result: Every module is a quotient of a free module). The next result is the dual result for $\mathbb{Z}$-modules: Every abelian group can be imbedded as a subgroup of an injective abelian group. We will prove this result for modules over any ring in Chapter 8 (see Theorem 8.104).

Corollary 7.75. *Every abelian group M can be imbedded as a subgroup of some injective abelian group.*

Proof. By Proposition 7.51, there is a free abelian group $F = \sum_i \mathbb{Z}_i$ with $M = F/K$ for some $K \subseteq F$. Now

$$M = F/K = \big(\sum_i \mathbb{Z}_i\big)/K \subseteq \big(\sum_i \mathbb{Q}_i\big)/K,$$

where we have merely imbedded each copy $\mathbb{Z}_i$ of $\mathbb{Z}$ into a copy $\mathbb{Q}_i$ of $\mathbb{Q}$. But Example 7.71 gives each $\mathbb{Q}_i$ divisible, hence gives $\sum_i \mathbb{Q}_i$ divisible, and hence gives divisibility of the quotient $(\sum_i \mathbb{Q}_i)/K$. By the Proposition, $(\sum_i \mathbb{Q}_i)/K$ is injective. •

Writing a module as a quotient of a free module is the essence of describing it by generators and relations. We may think of the corollary as dualizing this idea.

The next result gives a curious example of an injective module; we shall actually use it to prove an interesting result (see the remark on page 654 after the proof of the basis theorem).

Proposition 7.76. *Let R be a* PID*, let $a \in R$ be neither zero nor a unit, and let $J = (a)$. Then R/J is an injective R/J-module.*

Proof. By the correspondence theorem, every ideal in R/J has the form I/J for some ideal I in R containing J. Now $I = (b)$ for some $b \in I$, so that I/J is cyclic with generator $x = b + J$. Since $(a) \subseteq (b)$, we have $a = rb$ for some $r \in R$. We are going to use the Baer criterion, Theorem 7.68, to prove that R/J is injective.

Assume that $f : I/J \to R/J$ is an R/J-map, and write $f(b + J) = s + J$ for some $s \in R$. Since $r(b+J) = rb+J = a+J = 0$, we have $rf(b+J) = r(s+J) = rs+J = 0$, and so $rs \in J = (a)$. Hence, there is some $r' \in R$ with $rs = r'a = r'br$; canceling r gives $s = r'b$. Thus,

$$f(b + J) = s + J = r'b + J.$$

Define $h: R/J \to R/J$ to be multiplication by r'; that is, $h: u + J \mapsto r'u + J$. The displayed equation gives $h(b + J) = f(b + J)$, so that h does extend f. Therefore, R/J is injective. •

EXERCISES

7.40 Let M be a free R-module, where R is a domain. Prove that if $rm = 0$, where $r \in R$ and $m \in M$, then either $r = 0$ or $m = 0$. (This is false if R is not a domain.)

7.41 Use left exactness of Hom to prove that if G is an abelian group, then $\mathrm{Hom}_{\mathbb{Z}}(\mathbb{I}_n, G) \cong G[n]$, where $G[n] = \{g \in G : ng = 0\}$.

7.42 Prove that a group $G \in \mathrm{obj}(\mathbf{Groups})$ is a projective object if and only if G is a free group. (It is proved, in Exercise 10.3 on page 793, that the only injective object in **Groups** is $\{1\}$.)

7.43 If R is a domain but not a field, and if $Q = \text{Frac}(R)$, prove that

$$\text{Hom}_R(Q, R) = \{0\}.$$

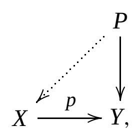
7.44 Prove that every left exact covariant functor $T\colon {}_R\textbf{Mod} \to \textbf{Ab}$ preserves pullbacks. Conclude that if B and C are submodules of a module A, then for every module M, we have

$$\text{Hom}_R(M, B \cap C) = \text{Hom}_R(M, B) \cap \text{Hom}_R(M, C).$$

Hint. Use pullback.

7.45 (i) Prove that a function is an epimorphism in **Sets** if and only if it is a surjection.

(ii) Prove that every object in **Sets** is projective, where an object P in a category is projective if a dotted arrow always exists for the diagram



where p is an epimorphism.

Hint. Use the axiom of choice.

7.46 Given a set X, prove that there exists a free R-module F with a basis B for which there is a bijection $\varphi\colon B \to X$.

7.47 (i) Prove that every vector space V over a field k is a free k-module.

(ii) Prove that a subset B of V is a basis of V considered as a vector space if and only if B is a basis of V considered as a free k-module.

7.48 Define G to be the abelian group having the presentation $(\mathcal{X}|\mathcal{R})$, where

$$\mathcal{X} = \{a, b_1, b_2, \ldots, b_n, \ldots\} \quad \text{and} \quad \mathcal{R} = \{2a, a - 2^n b_n, n \geq 1\}.$$

Thus, $G = F/K$, where F is the free abelian group with basis $\mathcal{X}$ and K is the subgroup $\langle \mathcal{R} \rangle$.

(i) Prove that $a + K \in G$ is nonzero.

(ii) Prove that $z = a + K$ satifies equations $z = 2^n y_n$, where $y_n \in G$ and $n \geq 1$, and that z is the unique such element of G.

(iii) Prove that there is an exact sequence $0 \to \langle a \rangle \to G \to \sum_{n \geq 1} \mathbb{I}_{2^n} \to 0$.

(iv) Prove that $\text{Hom}_{\mathbb{Z}}(\mathbb{Q}, G) = \{0\}$ by applying $\text{Hom}_{\mathbb{Z}}(\mathbb{Q}, \ )$ to the exact sequence in part (iii).

7.49 (i) If $\{P_i : i \in I\}$ is a family of projective R-modules, prove that their direct sum $\sum_{i \in I} P_i$ is also projective.

(ii) Prove that every direct summand of a projective module is projective.

7.50 Prove that every direct summand of an injective module is injective.

7.51 Give an example of two injective submodules of a module whose intersection is not injective.
Hint. Define abelian groups $A \cong \mathbb{Z}(p^\infty) \cong A'$:

$$A = (a_n, n \geq 0 | pa_0 = 0, pa_{n+1} = a_n) \quad \text{and} \quad A' = (a'_n, n \geq 0 | pa'_0 = 0, pa'_{n+1} = a'_n).$$

In $A \oplus A'$, define $E = A \oplus \{0\}$ and $E' = \langle \{(a_{n+1}, a'_n) : n \geq 0\} \rangle$.

7.52 (i) Prove that if a domain R is an injective R-module, then R is a field.

(ii) Let R be a domain that is not a field, and let M be an R-module that is both injective and projective. Prove that $M = \{0\}$.

(iii) Prove that $\mathbb{I}_6$ is simultaneously an injective and a projective module over itself.

7.53 (i) If R is a domain and I and J are nonzero ideals in R, prove that $I \cap J \neq \{0\}$.

(ii) Let R be a domain and let I be an ideal in R that is a free R-module; prove that I is a principal ideal.

7.54 Prove that an R-module E is injective if and only if, for every ideal I in R, every short exact sequence $0 \to E \to B \to I \to 0$ splits.

7.55 Prove the dual of Schanuel's lemma. Given exact sequences

$$0 \to M \xrightarrow{i} E \xrightarrow{p} Q \to 0 \quad \text{and} \quad 0 \to M \xrightarrow{i'} E' \xrightarrow{p'} Q' \to 0,$$

where E and E' are injective, then there is an isomorphism

$$Q \oplus E' \cong Q' \oplus E.$$

7.56 (i) Prove that every vector space over a field k is an injective k-module.

(ii) Prove that if $0 \to U \to V \to W \to 0$ is an exact sequence of vector spaces, then the corresponding sequence of dual spaces $0 \to W^* \to V^* \to U^* \to 0$ is also exact.

7.57 (***Pontrjagin Duality***) If G is an abelian group, its ***Pontrjagin dual*** is the group

$$G^* = \operatorname{Hom}_{\mathbb{Z}}(G, \mathbb{Q}/\mathbb{Z}).$$

(Pontrjagin duality extends to locally compact abelian topological groups, and the dual consists of all continuous homomorphisms into the circle group.)

(i) Prove that if G is an abelian group and $a \in G$ is nonzero, then there is a homomorphism $f : G \to \mathbb{Q}/\mathbb{Z}$ with $f(a) \neq 0$.

(ii) Prove that $\mathbb{Q}/\mathbb{Z}$ is an injective abelian group.

(iii) Prove that if $0 \to A \to G \to B \to 0$ is an exact sequence of abelian groups, then so is $0 \to B^* \to G^* \to A^* \to 0$.

(iv) If $G \cong \mathbb{I}_n$, prove that $G^* \cong G$.

(v) If G is a finite abelian group, prove that $G^* \cong G$.

(vi) Prove that if G is a finite abelian group, and if G/H is a quotient group of G, then G/H is isomorphic to a subgroup of G. [The analogous statement for nonabelian groups is false: If $\mathbf{Q}$ is the group of quaternions, then $\mathbf{Q}/Z(\mathbf{Q}) \cong \mathbf{V}$, where $\mathbf{V}$ is the four-group; but $\mathbf{Q}$ has only one element of order 2 while $\mathbf{V}$ has three elements of order 2. This exercise is also false for infinite abelian groups: Since $\mathbb{Z}$ has no element of order 2, it has no subgroup isomorphic to $\mathbb{Z}/2\mathbb{Z} \cong \mathbb{I}_2$.]

7.5 Grothendieck Groups

A. Grothendieck introduced abelian groups to help study projective modules. The reader may regard this section as a gentle introduction to algebraic K-theory.

Definition. A category $\mathcal{C}$ is a $\star$-category if there is a commutative and associative binary operation $\star\colon \operatorname{obj}(\mathcal{C}) \times \operatorname{obj}(\mathcal{C}) \to \operatorname{obj}(\mathcal{C})$; that is,

(i) If $A \cong A'$ and $B \cong B'$, where $A, A', B, B' \in \operatorname{obj}(\mathcal{C})$, then $A \star B \cong A' \star B'$.

(ii) there is an equivalence $A \star B \cong B \star A$ for all $A, B \in \operatorname{obj}(\mathcal{C})$;

(iii) there is an equivalence $A \star (B \star C) \cong (A \star B) \star C$ for all $A, B, C \in \operatorname{obj}(\mathcal{C})$.

Any category having finite products or finite coproducts is a $\star$-category.

Definition. If $\mathcal{C}$ is a $\star$-category, define $|\operatorname{obj}(\mathcal{C})|$ to be the class of all isomorphism classes $|A|$ of objects in $\mathcal{C}$, where $|A| = \{B \in \operatorname{obj}(\mathcal{C}) : B \cong A\}$. If $\mathcal{F}(\mathcal{C})$ is the free abelian group with basis[15] $|\operatorname{obj}(\mathcal{C})|$ and $\mathcal{R}$ is the subgroup of $\mathcal{F}(\mathcal{C})$ generated by all elements of the form

$$|A \star B| - |A| - |B| \quad \text{where } A, B \in \operatorname{obj}(\mathcal{C}),$$

then the ***Grothendieck group*** $K_0(\mathcal{C})$ is the abelian group

$$K_0(\mathcal{C}) = \mathcal{F}(\mathcal{C})/\mathcal{R}.$$

(A characterization of $K_0(\mathcal{C})$ as a solution to a universal mapping problem is given in Exercise 7.58 on page 498.) For any object A in $\mathcal{C}$, we denote the coset $|A| + \mathcal{R}$ by $[A]$.

We remark that the Grothendieck group $K_0(\mathcal{C})$ can be defined more precisely: $\mathcal{C}$ should be a *symmetric monoidal category* (see Mac Lane, *Categories for the Working Mathematician*, pages 157–161).

Proposition 7.77. *Let $\mathcal{C}$ be a $\star$-category.*

(i) *If $x \in K_0(\mathcal{C})$, then $x = [A] - [B]$ for $A, B \in \operatorname{obj}(\mathcal{C})$.*

(ii) *If $A, B \in \operatorname{obj}(\mathcal{C})$, then $[A] = [B]$ in $K_0(\mathcal{C})$ if and only if there exists $C \in \operatorname{obj}(\mathcal{C})$ with $A \star C \cong B \star C$.*

Proof. (i) Since $K_0(\mathcal{C})$ is generated by $|\operatorname{obj}(\mathcal{C})|$, we may write

$$x = \sum_{i=1}^{r} [A_i] - \sum_{j=1}^{s} [B_j],$$

(we allow objects A_i and B_j to be repeated). If we now define $A = A_1 \star \cdots \star A_r$, then

$$[A] = [A_1 \star \cdots \star A_r] = \sum_i [A_i].$$

Similarly, define $B = B_1 \star \cdots \star B_s$. It is now clear that $x = [A] - [B]$.

[15]There is a minor set-theoretic problem here, for a basis of a free abelian group must be a set and not a proper class. This problem is usually avoided by assuming that $\mathcal{C}$ is a ***small category***; that is, the class $\operatorname{obj}(\mathcal{C})$ is a set.

(ii) If $A \star C \cong B \star C$, then $[A \star C] = [B \star C]$ in $K_0(\mathcal{C})$. Hence, $[A] + [C] = [B] + [C]$, and the cancellation law in the abelian group $K_0(\mathcal{C})$ gives $[A] = [B]$.

Conversely, if $[A] = [B]$, then $|B| - |A| \in \mathcal{R}$ and there is an equation in $\mathcal{F}(\mathcal{C})$:

$$|B| - |A| = \sum_i m_i(|X_i \star Y_i| - |X_i| - |Y_i|) - \sum_j n_j(|U_j \star V_j| - |U_j| - |V_j|),$$

where the coefficients m_i and n_j are positive integers, and the X, Y, U, and V are objects in $\mathcal{C}$. Transposing to eliminate negative coefficients,

$$|A| + \sum_i m_i|X_i \star Y_i| + \sum_j n_j(|U_j| + |V_j|) = |B| + \sum_i m_i(|X_i| + |Y_i|) + \sum_j n_j|U_j \star V_j|.$$

This is an equation in a free abelian group, where expressions in terms of a basis are unique. Therefore, $\{A, X_i \star Y_i, U_j, V_j\}$, the set of objects, with multiplicities, on the left-hand side, coincides with $\{B, U_j \star V_j, X_i, Y_i\}$, the set of objects, with multiplicities, on the right-hand side. Since products are commutative and associative, there is an equivalence in $\mathcal{C}$:

$$A \star \big(\bigstar_i\, m_i(X_i \star Y_i)\big) \star \big(\bigstar_j\, n_j(U_j \star V_j)\big) \cong B \star \big(\bigstar_i\, m_i(X_i \star Y_i)\big) \star \big(\bigstar_j\, n_j(U_j \star V_j)\big).$$

An inspection of terms shows that

$$\big(\bigstar_i\, m_i(X_i \star Y_i)\big) \star \big(\bigstar_j\, n_j(U_j \star V_j)\big) \cong \big(\bigstar_i\, m_i(X_i \star Y_i)\big) \star \big(\bigstar_j\, n_j(U_j \star V_j)\big).$$

If we denote this last object by C, then $A \star C \cong B \star C$. •

Definition. Let R be a commutative ring, and let $\mathcal{C}$ be a subcategory of ${}_R\mathbf{Mod}$. Two R-modules A and B are called ***stably isomorphic in*** $\mathcal{C}$ if there exists a module $C \in \operatorname{obj}(\mathcal{C})$ with $A \oplus C \cong B \oplus C$.

With this terminology, Proposition 7.77 says that two modules determine the same element of a Grothendieck group if and only if they are stably isomorphic. It is clear that isomorphic modules are stably isomorphic; the next example shows that the converse need not hold.

Example 7.78.

(i) If **Ab** is the category of all finite abelian groups, then Exercise 5.10 on page 268 shows that two finite abelian groups are stably isomorphic in **Ab** if and only if they are isomorphic.

(ii) If R is a commutative ring and F is a free R-module of infinite rank, then

$$R \oplus F \cong R \oplus R \oplus F.$$

Thus, R and $R \oplus R$ are nonisomorphic modules that are stably isomorphic in ${}_R\mathbf{Mod}$. Because of examples of this type, we usually restrict ourselves to subcategories $\mathcal{C}$ of ${}_R\mathbf{Mod}$ consisting of finitely generated modules.

(iii) Here is an example, due to R. G. Swan, in which stable isomorphism of finitely generated projective modules does not imply isomorphism.

Let $R = \mathbb{R}[x_1, \ldots, x_n]/(1 - \sum_i x_i^2)$ [the coordinate ring of the real $(n-1)$-sphere]. Regard R^n as $n \times 1$ column vectors, and let $X = (\overline{x}_1, \ldots, \overline{x}_n)^t \in R^n$, where bar denotes coset mod $(1 - \sum_i x_i^2)$ in R. Define $\lambda\colon R \to R^n$ by $\lambda\colon r \mapsto rX$, and define $\varphi\colon R^n \to R$ by $\varphi(Y) = X^t Y$. Note that the composite $\varphi\lambda\colon R \to R$ is the identity, for

$$\varphi\lambda(r) = \varphi(rX) = X^t r X = r,$$

because $X^t X = \sum_i \overline{x}_i^2 = 1$. It follows that the exact sequence

$$0 \to R \xrightarrow{\lambda} R^n \xrightarrow{\text{nat}} R^n/\operatorname{im}\lambda \to 0$$

splits. Thus, if $P = R^n/\operatorname{im}\lambda$, then

$$R \oplus R^{n-1} \cong R^n \cong R \oplus P,$$

and P is stably isomorphic to the free R-module R^{n-1} (of course, P is a projective R-module). Using topology, Swan proved that P is a free R-module if and only if $n = 1, 2, 4$ or 8. If $n = 3$, for example, then P is not isomorphic to R^{n-1}. ◂

Proposition 7.79. *If $\mathcal{C}$ is the category of all finite abelian groups, then $K_0(\mathcal{C})$ is a free abelian group with a basis $\mathcal{B}$ consisting of all the cyclic primary groups.*

Proof. By the basis theorem, each finite abelian group $A \cong \sum_i C_i$, where each C_i is a cyclic primary group. Thus, $[A] = \sum_i [C_i]$ in $K_0(\mathcal{C})$. Since every element in $K_0(\mathcal{C})$ is equal to $[A] - [B]$, for finite abelian groups A and B, it follows that $\mathcal{B}$ generates $K_0(\mathcal{C})$. To see that $\mathcal{B}$ is a basis, suppose that $\sum_{i=1}^r m_i[C_i] - \sum_{j=1}^s n_j[C_j'] = 0$, where m_i and n_j are positive integers. Then $\sum_i [m_i C_i] = \sum_j [n_j C_j']$, where $m_i C_i$ is the direct sum of m_i copies of C_i, and so $[\sum_i m_i C_i] = [\sum_j n_j C_j']$. Therefore, $\sum_i m_i C_i$ and $\sum_j n_j C_j'$ are stably isomorphic in $\mathcal{C}$. By Example 7.78(i), $\sum_{i=1}^r m_i C_i \cong \sum_{j=1}^s n_j C_j'$. Finally, the fundamental theorem of finite abelian groups applies to give $r = s$, a permutation $\sigma \in S_r$ with $C_{\sigma(i)}' \cong C_i$ and $m_i = n_{\sigma(i)}$ for all i. Therefore, $\mathcal{B}$ is a basis of $K_0(\mathcal{C})$. •

Definition. If R is a commutative ring, then the subcategory $\mathbf{Pr}(R)$ of all finitely generated projective R-modules is a $\star$-category (for the direct sum of two such modules is again finitely generated projective). We usually denote $K_0(\mathbf{Pr}(R))$ by $K_0(R)$ in this case.

Example 7.80.
We now show that $K_0(R) \cong \mathbb{Z}$ if R is a commutative ring for which every finitely generated projective R-module is free. It is clear that $K_0(R)$ is generated by $[R]$, so that $K_0(R)$ is cyclic. Define $r\colon \text{obj}(\mathbf{Pr}(R)) \to \mathbb{Z}$ by $r(F) = \text{rank}(F)$, where F is a finitely generated free R-module. Since $r(F \oplus F') = r(F) + r(F')$, Exercise 7.58 on page 498 shows that there is a homomorphism $\widetilde{r}\colon K_0(R) \to \mathbb{Z}$ with $\widetilde{r}([F]) = \text{rank}(F)$ for every finitely generated free F. Since $K_0(R)$ is cyclic, $\widetilde{r}$ is an isomorphism. ◂

If $\mathcal{C}$ is a category of modules, there is another Grothendieck group $K'(\mathcal{C})$ we can define.

Definition. If $\mathcal{C}$ is a category of modules, define $\mathcal{F}(\mathcal{C})|$ to be the free abelian group with basis $|\text{obj}(\mathcal{C})$, and $\mathcal{R}'$ to be the subgroup of $\mathcal{F}(\mathcal{C})$ generated by all elements of the form

$$|B| - |A| - |C| \quad \text{if there is an exact sequence} \quad 0 \to A \to B \to C \to 0.$$

The ***Grothendieck group*** $K'(\mathcal{C})$ is the abelian group

$$K'(\mathcal{C}) = \mathcal{F}(\mathcal{C})/\mathcal{R}';$$

that is, $K'(\mathcal{C})$ is the abelian group with generators $|\text{obj}(\mathcal{C})|$ and relations $\mathcal{R}'$. For any module $A \in \text{obj}(\mathcal{C})$, we denote the coset $|A| + \mathcal{R}'$ by (A).

Example 7.81.
If R is a domain and $a \in R$ is neither 0 nor a unit, there is an exact sequence

$$0 \to R \xrightarrow{\mu_a} R \to R/Ra \to 0,$$

where $\mu_a\colon r \mapsto ar$. Thus, there is an equation in $K'(\mathcal{C})$:

$$(R) = (R) + (R/Ra).$$

Hence, $(R/Ra) = 0$. ◂

The next proposition uses the observation that the two notions of Grothendieck group—$K_0(R) = K_0(\mathbf{Pr}(R))$ and $K'(\mathbf{Pr}(R))$—coincide. The reason is that there is an exact sequence $0 \to A \to A \oplus C \to C \to 0$, so that $(A \oplus C) = (A) + (C)$ in $K'(\mathcal{C})$.

Proposition 7.82. *If R is a commutative ring and $\mathcal{C}$ is the category of all finitely generated R-modules, then there is a homomorphism*

$$\varepsilon\colon K_0(R) \to K'(\mathcal{C})$$

with $\varepsilon\colon [P] \mapsto (P)$ for every projective R-module P.

Proof. Since every short exact sequence of projective modules splits, the relations defining $K_0(R) = K_0(\mathbf{Pr}(R))$ are the same as those defining $K'(\mathbf{Pr}(R))$. Hence, the inclusion map $\mathcal{F}(\mathbf{Pr}(R)) \to \mathcal{F}(\mathcal{C})$ induces a well-defined homomorphism. •

Proposition 7.83. *Let R be a commutative ring and let $\mathcal{C}$ be the category of all finitely generated R-modules. If $M \in \operatorname{obj}(\mathcal{C})$ and*

$$M = M_0 \supseteq M_1 \supseteq M_2 \supseteq \cdots \supseteq M_n = \{0\}$$

has factor modules $Q_i = M_{i-1}/M_i$, then

$$(M) = (Q_1) + \cdots + (Q_n) \text{ in } K'(\mathcal{C}).$$

Proof. Since $Q_i = M_{i-1}/M_i$, there is a short exact sequence

$$0 \to M_i \to M_{i-1} \to Q_i \to 0,$$

so that $(Q_i) = (M_{i-1}) - (M_i)$ in $K'(\mathcal{C})$. We now have a telescoping sum:

$$\sum_{i=1}^{n}(Q_i) = \sum_{i=1}^{n}\big[(M_{i-1}) - (M_i)\big] = (M_0) - (M_n) = (M). \quad \bullet$$

The next obvious question is how to detect when an element in $K'(\mathcal{C})$ is zero.

Proposition 7.84. *Let R be a commutative ring and let $\mathcal{C}$ be the category of all finitely generated R-modules. If $A, B \in \operatorname{obj}(\mathcal{C})$, then $(A) = (B)$ in $K'(\mathcal{C})$ if and only if there are $C, U, V \in \operatorname{obj}(\mathcal{C})$ and exact sequences*

$$0 \to U \to A \oplus C \to V \to 0 \quad \textit{and} \quad 0 \to U \to B \oplus C \to V \to 0.$$

Proof. If there exist modules C, U, and V as in the statement, then

$$(A \oplus C) = (U) + (V) = (B \oplus C).$$

But exactness of $0 \to A \to A \oplus C \to C \to 0$ gives $(A \oplus C) = (A) + (C)$. Similarly, $(B \oplus C) = (B) + (C)$, so that $(A) + (C) = (B) + (C)$ and $(A) = (B)$.

Conversely, if $(A) = (B)$, then $|A| - |B| \in \mathcal{R}'$. As in the proof of Proposition 7.77, there is an equation in $\mathcal{F}(\mathcal{C})$:

$$|A| + \sum |X_i| + \sum(|Y_j'| + |Y_j''|) = |B| + \sum(|X_i'| + |X_i''|) + \sum |Y_j|,$$

where $0 \to X_i' \to X_i \to X_i'' \to 0$ and $0 \to Y_j' \to Y_j \to Y_j'' \to 0$ are exact sequences. Define

$$C = A \oplus \sum X_i \oplus \sum(Y_j' \oplus Y_j'').$$

Setting X' to be the direct sum of the X_i', X to be the direct sum of the X_i, and so forth, the argument in the proof of Proposition 7.77 gives

$$A \oplus X \oplus Y' \oplus Y'' \cong B \oplus X' \oplus X'' \oplus Y.$$

By Exercise 7.12 on page 440, this isomorphism gives rise to exact sequences

$$0 \to X' \oplus Y'' \to X \oplus Y'' \to X'' \to 0,$$

$$0 \to X' \oplus Y'' \to (X \oplus Y'') \oplus Y' \to X'' \oplus Y' \to 0,$$

and

$$0 \to X' \oplus Y'' \to A \oplus (X \oplus Y' \oplus Y'') \to A \oplus (X'' \oplus Y') \to 0.$$

The middle module is C. Applying Exercise 7.12 once again, there is an exact sequence

$$0 \to X' \oplus Y' \to B \oplus C \to B \oplus (A \oplus X'' \oplus Y'') \to 0.$$

Define $U = X' \oplus Y'$ and $V = B \oplus A \oplus X'' \oplus Y''$; with this notation, the last exact sequence is

$$0 \to U \to B \oplus C \to V \to 0.$$

Similar manipulation yields an exact sequence $0 \to U \to A \oplus C \to V \to 0$. •

In Chapter 8, we will prove a module version of Theorem 5.52, the Jordan–Hölder theorem. For now, we merely give a definition.

Definition. A sequence in a category $\mathcal{C}$ of modules,

$$M = M_0 \supseteq M_1 \supseteq M_2 \supseteq \cdots \supseteq M_n = \{0\},$$

is called a ***composition series*** of M if each of its factor modules $Q_i = M_{i-1}/M_i$ is a simple module in $\operatorname{obj}(\mathcal{C})$. We say that a category $\mathcal{C}$ of modules is a ***Jordan–Hölder category*** if:

(i) Each object M has a composition series;

(ii) For every two composition series

$$M = M_0 \supseteq M_1 \supseteq M_2 \supseteq \cdots \supseteq M_n = \{0\}$$

and

$$M = M'_0 \supseteq M'_1 \supseteq M'_2 \supseteq \cdots \supseteq M'_m = \{0\},$$

we have $m = n$ and a permutation $\sigma \in S_n$ such that $Q'_j \cong Q_{\sigma j}$ for all j, where $Q_i = M_{i-1}/M_i$ and $Q'_j = M'_{j-1}/M'_j$.

Define the ***length*** $\ell(M)$ of a module M in a Jordan–Hölder category to be the number n of terms in a composition series. If the simple factor modules of a composition series are $Q_1, \ldots, Q_n$, we define

$$\operatorname{jh}(M) = Q_1 \oplus \cdots \oplus Q_n.$$

A composition series may have several isomorphic factor modules, and $\operatorname{jh}(M)$ records their multiplicity.

Lemma 7.85. *Let $\mathcal{C}$ be a Jordan–Hölder category, and let $Q_1, \ldots, Q_n$, $Q'_1, \ldots, Q'_m$ be simple modules in* $\text{obj}(\mathcal{C})$.

(i) *If*

$$Q_1 \oplus \cdots \oplus Q_n \cong Q'_1 \oplus \cdots \oplus Q'_m,$$

then $m = n$ and there is a permutation $\sigma \in S_n$ such that $Q'_j \cong Q_{\sigma j}$ for all j, where $Q_i = M_{i-1}/M_i$ and $Q'_j = M'_{j-1}/M'_j$.

(ii) *If M and M' are modules in* $\text{obj}(\mathcal{C})$, *and if there is a simple module $S \in$* $\text{obj}(\mathcal{C})$ *with*

$$S \oplus \text{jh}(M) \cong S \oplus \text{jh}(M'),$$

then $\text{jh}(M) \cong \text{jh}(M')$.

Proof. (i) Now

$$Q_1 \oplus \cdots \oplus Q_n \supseteq Q_2 \oplus \cdots \oplus Q_n \supseteq Q_3 \oplus \cdots \oplus Q_n \supseteq \cdots$$

is a composition series with factor modules $Q_1, \ldots, Q_n$; similarly, the isomorphic module $Q'_1 \oplus \cdots \oplus Q'_m$ has a composition series with factor modules $Q'_1, \ldots, Q'_m$. The result follows from $\mathcal{C}$ being a Jordan–Hölder category.

(ii) This result follows from part (i) because S is simple. •

Lemma 7.86. *If $0 \to A \to B \to C \to 0$ is an exact sequence in a Jordan–Hölder category, then*

$$\text{jh}(B) \cong \text{jh}(A) \oplus \text{jh}(C).$$

Proof. The proof is by induction on the length $\ell(C)$. Let $A = A_0 \supseteq A_1 \supseteq \cdots \supseteq A_n = \{0\}$ be a composition series for A with factor modules $Q_1, \ldots, Q_n$. If $\ell(C) = 1$, then C is simple, and so

$$B \supseteq A \supseteq A_1 \supseteq \cdots \supseteq A_n = \{0\}$$

is a composition series for B with factor modules $C, Q_1, \ldots, Q_n$. Therefore,

$$\text{jh}(B) = C \oplus Q_1 \oplus \cdots \oplus Q_n = \text{jh}(C) \oplus \text{jh}(A).$$

For the inductive step, let $\ell(C) > 1$. Choose a maximal submodule C_1 of C (which exists because C has a composition series). If $\nu \colon B \to C$ is the given surjection, define $B_1 = \nu^{-1}(C_1)$. There is a commutative diagram (with vertical arrows inclusions)

$$\begin{array}{ccccccccc}
0 & \longrightarrow & A & \longrightarrow & B & \xrightarrow{\nu} & C & \longrightarrow & 0 \\
 & & \uparrow & & \uparrow & & \uparrow & & \\
0 & \longrightarrow & A & \longrightarrow & B_1 & \longrightarrow & C_1 & \longrightarrow & 0
\end{array}$$

Since C_1 is a maximal submodule of C, the quotient module

$$C'' = C/C_1$$

is simple. Note that $B/B_1 \cong (B/A)/(B_1/A) \cong C/C_1 = C''$. By the base step, we have

$$\text{jh}(C) = C'' \oplus \text{jh}(C_1) \quad \text{and} \quad \text{jh}(B) = C'' \oplus \text{jh}(B_1).$$

By the inductive hypothesis,

$$\text{jh}(B_1) = \text{jh}(A) \oplus \text{jh}(C_1).$$

Therefore,

$$\begin{aligned} \text{jh}(B) &= C'' \oplus \text{jh}(B_1) \\ &\cong C'' \oplus \text{jh}(A) \oplus \text{jh}(C_1) \\ &\cong \text{jh}(A) \oplus C'' \oplus \text{jh}(C_1) \\ &\cong \text{jh}(A) \oplus \text{jh}(C). \quad \bullet \end{aligned}$$

Theorem 7.87. *Let $\mathcal{C}$ be a category of modules in which every module $M \in \text{obj}(\mathcal{C})$ has a composition series. Then $\mathcal{C}$ is a Jordan–Hölder category if and only if $K'(\mathcal{C})$ is a free abelian group with basis the set $\mathcal{B}'$ of all (S) as S varies over all nonisomorphic simple modules in* $\text{obj}(\mathcal{C})$.

Proof. Assume that $K'(\mathcal{C})$ is free abelian with basis $\mathcal{B}'$. Since 0 is not a member of a basis, we have $(S) \neq 0$ for every simple module S; moreover, if $S \ncong S'$, then $(S) \neq (S')$, for a basis repeats no elements. Let $M \in \text{obj}(\mathcal{C})$, and let $Q_1, \ldots, Q_n$ and $Q'_1, \ldots, Q'_m$ be simple modules arising, respectively, as factor modules of two composition series of M. By Proposition 7.83, we have

$$(Q_1) + \cdots + (Q_n) = (M) = (Q'_1) + \cdots + (Q'_m).$$

Uniqueness of expression in terms of the basis $\mathcal{B}'$ says, for each Q'_j, that there exists Q_i with $(Q_i) = (Q'_j)$; in fact, the number of any (Q_i) on the left-hand side is equal to the number of copies of (Q'_j) on the right-hand side. Therefore, $\mathcal{C}$ is a Jordan–Hölder category.

Conversely, assume that the Jordan–Hölder theorem holds for $\mathcal{C}$. Since every $M \in \text{obj}(\mathcal{C})$ has a composition series, Proposition 7.83 shows that $\mathcal{B}'$ generates $K'(\mathcal{C})$. Let S be a simple module in $\text{obj}(\mathcal{C})$. If $(S) = (T)$, then Proposition 7.84 says there are $C, U, V \in \text{obj}(\mathcal{C})$ and exact sequences $0 \to U \to S \oplus C \to V \to 0$ and $0 \to U \to T \oplus C \to V \to 0$. Lemma 7.86 gives

$$\text{jh}(S) \oplus \text{jh}(C) \cong \text{jh}(U) \oplus \text{jh}(V) \cong \text{jh}(T) \oplus \text{jh}(C).$$

By Lemma 7.85, we may cancel the simple summands one by one until we are left with $S \cong T$, a contradiction. A similar argument shows that if S is a simple module, then

$(S) \neq 0$. Finally, let us show that every element in $K'(\mathcal{C})$ has a unique expression as a linear combination of elements in $\mathcal{B}'$. Suppose there are positive integers m_i and n_j so that

$$\sum_i m_i(S_i) - \sum_j n_j(T_j) = 0, \tag{1}$$

where the S_i and T_j are simple modules in $\text{obj}(\mathcal{C})$ and $S_i \not\cong T_j$ for all i, j. If we denote the direct sum of m_i copies of S_i by $m_i S_i$, then Eq. (1) gives

$$\big(\sum_i m_i S_i\big) = \big(\sum_j n_j T_j\big).$$

By Proposition 7.84, there are modules C, U, V and exact sequences

$$0 \to U \to C \oplus \sum_i m_i S_i \to V \to 0 \quad \text{and} \quad 0 \to U \to C \oplus \sum_j n_j T_j \to V \to 0,$$

and Lemma 7.86 gives

$$\text{jh}\big(\sum_i m_i S_i\big) \cong \text{jh}\big(\sum_j n_j T_j\big).$$

By Lemma 7.85, some S_i occurs on the right-hand side, contradicting $S_i \not\cong T_j$ for all i, j. Therefore, Eq. (1) cannot occur. •

Remark. A module M is called ***indecomposable*** if there do not exist nonzero modules A and B with $M \cong A \oplus B$. We say that a category $\mathcal{C}$ of modules is a ***Krull–Schmidt category*** if:

(i) Each module in $\text{obj}(\mathcal{C})$ is isomorphic to a finite direct sum of indecomposable modules in $\text{obj}(\mathcal{C})$;

(ii) If
$$D_1 \oplus \cdots \oplus D_n \cong D'_1 \oplus \cdots \oplus D'_m,$$
where all the summands are indecomposable, then $m = n$ and there is a permutation $\sigma \in S_n$ with $D'_j \cong D_{\sigma j}$ for all j.

There is a theorem analogous to Theorem 7.87 saying that a category $\mathcal{C}$ of modules is a Krull-Schmidt category if and only if $K_0(\mathcal{C})$ is a free abelian group with basis consisting of all $[D]$ as D varies over all nonisomorphic indecomposable modules in $\text{obj}(\mathcal{C})$. ◂

Compare the next corollary with Proposition 7.79.

Corollary 7.88. *If $\mathcal{C}$ is the category of all finite abelian groups, then $K'(\mathcal{C})$ is the free abelian group with generators all (S), where S is a cyclic group of prime order p.*

Proof. By Theorem 5.52, the category of all finite abelian groups is a Jordan–Hölder category, and the simple $\mathbb{Z}$-modules are the abelian groups $\mathbb{I}_p$ for primes p. •

H. Bass defined higher groups K_1 and K_2, proved that there is an exact sequence relating these to K_0, and showed how these groups can be used to study projective modules (see Milnor, *Introduction to Algebraic K-Theory*). D. Quillen constructed an infinite sequence $K_n(\mathcal{C})$ of abelian groups by associating a topological space $X(\mathcal{C})$ to certain categories $\mathcal{C}$. He then defined $K_n(\mathcal{C}) = \pi_{n+1}(X(\mathcal{C}))$ for all $n \geq 0$, the homotopy groups of this space, and he proved that his K_n coincide with those of Bass for $n = 0, 1, 2$ (see Rosenberg, *Algebraic K-Theory and Its Applications*).

EXERCISES

7.58 Let $\mathcal{C}$ be a $\star$-category. Prove that $K_0(\mathcal{C})$ solves the following universal mapping problem.

$$\begin{array}{ccc} \text{obj}(\mathcal{C}) & \xrightarrow{h} & K_0(\mathcal{C}), \\ {\scriptstyle f}\downarrow & \swarrow {\scriptstyle \widetilde{f}} & \\ G & & \end{array}$$

where G is any abelian group. If $h\colon \text{obj}(\mathcal{C}) \to K_0(\mathcal{C})$ is the function $A \mapsto [A]$, and if $f\colon \text{obj}(\mathcal{C}) \to G$ satisfies $f(A) = f(B)$ whenever $A \cong B$ and $f(A \star B) = f(A) + f(B)$, then there exists a unique homomorphism $\widetilde{f}\colon K_0(\mathcal{C}) \to G$ making the diagram commute.

7.59 Regard $\mathcal{C} = \mathbf{PO}(\mathbb{N})$ as a $\star$-category, where $m \star n = m + n$, and prove that $K_0(\mathcal{C}) \cong \mathbb{Z}$. (Thus, we have constructed the integers from the natural numbers. In a similar way, we can construct an abelian group G from a semigroup S, although we cannot expect that S is always imbedded in G.)

7.60 (i) If $\mathcal{C}$ is a category of modules having infinite direct sums of its objects, prove that $K_0(\mathcal{C}) = \{0\}$.

(ii) (***Eilenberg***) Prove that if P is a projective R-module (over some commutative ring R), then there exists a free R-module Q with $P \oplus Q$ a free R-module. Conclude that $K_0(\mathcal{C}) = \{0\}$ for $\mathcal{C}$ the category of countably generated projective R-modules.

Hint. Q need not be finitely generated.

7.61 Prove that $K_0(\mathbb{I}_6) \cong \mathbb{Z} \oplus \mathbb{Z}$.

7.62 Let $\mathcal{C}$ and $\mathcal{C}'$ be $\star$-categories, and let $F\colon \mathcal{C} \to \mathcal{C}'$ be a $\star$-preserving functor; that is, $F(A \star B) \cong F(A) \star F(B)$. Prove that F induces a homomorphism $K_0(\mathcal{C}) \to K_0(\mathcal{C}')$ by $[A] \mapsto [FA]$.

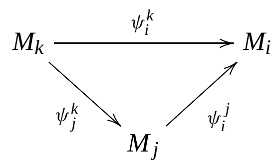
7.63 If $\mathcal{C}$ is a category of modules, prove that every element in $K'(\mathcal{C})$ has the form $(A) - (B)$ for modules A and B in $\text{obj}(\mathcal{C})$.

7.64 Let $\mathcal{C}$ be a category having short exact sequences. Prove that there is a surjection $K_0(\mathcal{C}) \to K'(\mathcal{C})$.

7.6 LIMITS

There are two more general constructions, one generalizing pullbacks and intersections, the other generalizing pushouts and unions; both involve a family of modules $\{M_i : i \in I\}$ whose index set I is a partially ordered set.

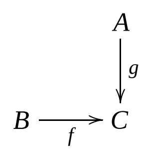
Definition. Let I be a partially ordered set. An ***inverse system of*** R***-modules*** over I is an ordered pair $\{M_i, \psi_i^j\}$ consisting of an indexed family of modules $\{M_i : i \in I\}$ together with a family of morphisms $\{\psi_i^j : M_j \to M_i\}$ for $i \preceq j$, such that $\psi_i^i = 1_{M_i}$ for all i and such that the following diagram commutes whenever $i \preceq j \preceq k$:



In Example 7.25(v), we saw that a partially ordered set I defines a category $\mathbf{PO}(I)$: The objects of $\mathbf{PO}(I)$ are the elements of I and $\mathrm{Hom}(i, j)$ is empty when $i \not\preceq j$ while it contains exactly one element, κ_j^i, whenever $i \preceq j$. If we define $F(i) = M_i$ and $F(\kappa_j^i) = \psi_i^j$, then it is easy to see that $\{M_i, \psi_i^j\}$ is an inverse system if and only if $F : \mathbf{PO}(I) \to {}_R\mathbf{Mod}$ is a contravariant functor. We now see that inverse systems involving objects and morphisms in any category $\mathcal{C}$ can be defined: Every contravariant functor $F : \mathbf{PO}(I) \to \mathcal{C}$ yields one. For example, we can speak of inverse systems of commutative rings.

Example 7.89.

(i) If $I = \{1, 2, 3\}$ is the partially ordered set in which $1 \preceq 2$ and $1 \preceq 3$, then an inverse system over I is a diagram of the form



(ii) If $\mathcal{I}$ is a family of submodules of a module A, then it can be partially ordered under *reverse inclusion*; that is, $M \preceq M'$ in case $M \supseteq M'$. For $M \preceq M'$, the inclusion map $M' \to M$ is defined, and it is easy to see that the family of all $M \in \mathcal{I}$ with inclusion maps is an inverse system.

(iii) If I is equipped with the ***discrete*** partial order, that is, $i \preceq j$ if and only if $i = j$, then an inverse system over I is just an indexed family of modules.

(iv) If $\mathbb{N}$ is the natural numbers with the usual partial order, then an inverse system over $\mathbb{N}$ is a diagram

$$M_0 \leftarrow M_1 \leftarrow M_2 \leftarrow \cdots .$$

(v) If J is an ideal in a commutative ring R, then its nth power is defined by

$$J^n = \left\{ \sum a_1 \cdots a_n : a_i \in J \right\}.$$

Each J^n is an ideal and there is a decreasing sequence

$$R \supseteq J \supseteq J^2 \supseteq J^3 \supseteq \cdots.$$

If A is an R-module, there is a sequence of submodules

$$A \supseteq JA \supseteq J^2 A \supseteq J^3 A \supseteq \cdots.$$

If $m \geq n$, define $\psi_n^m \colon A/J^m A \to A/J^n A$ by

$$\psi_n^m \colon a + J^m A \mapsto a + J^n A$$

(these maps are well-defined, for $m \geq n$ implies $J^m A \subseteq J^n A$). It is easy to see that

$$\{A/J^n A, \psi_n^m\}$$

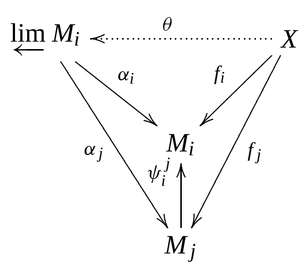
is an inverse system over $\mathbb{N}$.

(vi) Let G be a group and let $\mathcal{N}$ be the family of all the normal subgroups N of G having finite index partially ordered by reverse inclusion. If $N \preceq N'$ in $\mathcal{N}$, then $N' \leq N$; define $\psi_N^{N'} \colon G/N' \to G/N$ by $gN' \mapsto gN$. It is easy to see that the family of all such quotients together with the maps $\psi_N^{N'}$ form an inverse system over $\mathcal{N}$. ◂

Definition. Let I be a partially ordered set, and let $\{M_i, \psi_i^j\}$ be an inverse system of R-modules over I. The ***inverse limit*** (also called ***projective limit*** or ***limit***) is an R-module $\varprojlim M_i$ and a family of R-maps $\{\alpha_i : \varprojlim M_i \to M_i : i \in I\}$, such that

(i) $\psi_i^j \alpha_j = \alpha_i$ whenever $i \preceq j$;

(ii) for every module X having maps $f_i \colon X \to M_i$ satisfying $\psi_i^j f_j = f_i$ for all $i \preceq j$, there exists a unique map $\theta \colon X \to \varprojlim M_i$ making the following diagram commute:



The notation $\varprojlim M_i$ for an inverse limit is deficient in that it does not display the maps of the corresponding inverse system (and $\varprojlim M_i$ does depend on them). However, this is standard practice.

As with any object defined as a solution to a universal mapping problem, the inverse limit of an inverse system is unique (to isomorphism) if it exists.

Proposition 7.90. *The inverse limit of any inverse system $\{M_i, \psi_i^j\}$ of R-modules over a partially ordered index set I exists.*

Proof. Define

$$L = \{(m_i) \in \prod M_i : m_i = \psi_i^j(m_j) \text{ whenever } i \preceq j\};$$

it is easy to check that L is a submodule of $\prod_i M_i$. If p_i is the projection of the product to M_i, define $\alpha_i : L \to M_i$ to be the restriction $p_i|L$. It is clear that $\psi_i^j \alpha_j = \alpha_i$.

Assume that X is a module having maps $f_i : X \to M_i$ satisfying $\psi_i^j f_j = f_i$ for all $i \preceq j$. Define $\theta : X \to \prod M_i$ by

$$\theta(x) = (f_i(x)).$$

That $\operatorname{im}\theta \subseteq L$ follows from the given equation $\psi_i^j f_j = f_i$ for all $i \preceq j$. Also, θ makes the diagram commute: $\alpha_i\theta : x \mapsto (f_i(x)) \mapsto f_i(x)$. Finally, θ is the unique map $X \to L$ making the diagram commute for all $i \preceq j$. If $\varphi : X \to L$, then $\varphi(x) = (m_i)$ and $\alpha_i\varphi(x) = m_i$. Thus, if φ satisfies $\alpha_i\varphi(x) = f_i(x)$ for all i and all x, then $m_i = f_i(x)$, and so $\varphi = \theta$. We conclude that $L \cong \varprojlim M_i$. •

Inverse limits in categories other than module categories may exist; for example, inverse limits of commutative rings exist, as do inverse limits of groups or of topological spaces.

The reader should supply verifications of the following assertions in which we describe the inverse limit of each of the inverse systems in Example 7.89.

Example 7.91.

(i) If I is the partially ordered set $\{1, 2, 3\}$ with $1 \succeq 3$ and $2 \succeq 3$, then an inverse system is a diagram

$$\begin{array}{ccc} & & A \\ & & \downarrow{\scriptstyle g} \\ B & \xrightarrow[f]{} & C \end{array}$$

and the inverse limit is the pullback.

(ii) We have seen that the intersection of two submodules of a module is a special case of pullback. Suppose now that $\mathcal{I}$ is a family of submodules of a module A, so that $\mathcal{I}$ and inclusion maps is an inverse system, as in Example 7.89(ii). The inverse limit of this inverse system is $\bigcap_{M \in \mathcal{I}} M$.

(iii) If I is a discrete index set, then the inverse system $\{M_i : i \in I\}$ has the product $\prod_i M_i$ as its inverse limit. Indeed, this is just the diagrammatic definition of a product.

(iv) If J is an ideal in a commutative ring R and M is an R-module, then the inverse limit of $\{M/J^nM, \psi_n^m\}$ [in Example 7.89(v)] is usually called the ***J-adic completion*** of M; let us denote it by $\widehat{M}$. In order to understand the terminology, we give a rapid account of a corner of point-set topology.

Definition. A ***metric space*** is a set X equipped with a function $d\colon X \times X \to \mathbb{R}$, called a ***metric***, that satisfies the following axioms. For all $x, y, z \in X$,

(i) $d(x, y) \geq 0$ and $d(x, y) = 0$ if and only if $x = y$;

(ii) $d(x, y) = d(y, x)$;

(iii) (***Triangle inequality***) $d(x, y) \leq d(x, z) + d(z, y)$.

For example, $d(x, y) = |x - y|$ is a metric on $\mathbb{R}$. Given a metric space X, the usual definition of convergence of a sequence makes sense: A sequence (x_n) of points x_n in X ***converges*** to a ***limit*** $y \in X$ if, for every $\epsilon > 0$, there is N so that $d(x_n, y) < \epsilon$ whenever $n \geq N$; we denote (x_n) converging to y by

$$x_n \to y.$$

A difficulty with this definition is that we cannot tell if a sequence is convergent without knowing what its limit is. A sequence (x_n) is a ***Cauchy sequence*** if, for every $\epsilon > 0$, there is N so that $d(x_m, x_n) < \epsilon$ whenever $m, n \geq N$. The virtue of this condition on a sequence is that it involves only the terms of the sequence and not its limit. If $X = \mathbb{R}$, then a sequence is convergent if and only if it is a Cauchy sequence. In general metric spaces, however, we can prove that convergent sequences are Cauchy sequences, but the converse may be false. For example, if X consists of the positive real numbers, with the usual metric $|x - y|$, then the sequence $(1/n)$ is a Cauchy sequence, but it does not converge in X because $0 \notin X$.

Definition. A ***completion*** $\widehat{X}$ of a metric space X is a metric space with the following two properties:

(i) X is a *dense* subspace of $\widehat{X}$; that is, for every $\widehat{x} \in \widehat{X}$, there is a sequence (x_n) in X with $x_n \to \widehat{x}$;

(ii) every Cauchy sequence in $\widehat{X}$ converges to a limit in $\widehat{X}$.

It can be proved that any two completions of a metric space X are *isometric* (there is a bijection between them that preserves the metrics), and one way to prove existence of $\widehat{X}$ is to define its elements as equivalence classes of Cauchy sequences (x_n) in X, where we define $(x_n) \equiv (y_n)$ if $d(x_n, y_n) \to 0$.

Let us return to the inverse system $\{M/J^nM, \psi_n^m\}$. A sequence

$$(a_1 + JM, a_2 + J^2M, a_3 + J^3M, \ldots) \in \varprojlim (M/J^nM)$$

satisfies the condition $\psi_n^m(a_m + J^mM) = a_m + J^nM$ for all $m \geq n$, so that

$$a_m - a_n \in J^nM \quad \text{whenever } m \geq n.$$

This suggests the following metric on M in the (most important) special case when $\bigcap_{n=1}^{\infty} J^nM = \{0\}$. If $x \in M$ and $x \neq 0$, then there is i with $x \in J^iM$ and $x \notin J^{i+1}M$; define $\|x\| = 2^{-i}$; define $\|0\| = 0$. It is a routine calculation to see that $d(x, y) = \|x - y\|$

is a metric on M (without the intersection condition, $\|x\|$ would not be defined for a nonzero $x \in \bigcap_{n=1}^{\infty} J^n M$). Moreover, if a sequence (a_n) in M is a Cauchy sequence, then $(a_1 + JM, a_2 + J^2M, a_3 + J^3M, \ldots) \in \varprojlim M/J^n M$, and conversely.

In particular, when $M = \mathbb{Z}$ and $J = (p)$, where p is a prime, then the completion $\mathbb{Z}_p$ is called the ring of ***p*-adic integers**. It turns out that $\mathbb{Z}_p$ is a domain, and $\mathbb{Q}_p = \text{Frac}(\mathbb{Z}_p)$ is called the field of ***p*-adic numbers**.

(v) We have seen, in Example 7.89(vi), that the family $\mathcal{N}$ of all normal subgroups of finite index in a group G forms an inverse system; the inverse limit of this system, $\varprojlim G/N$, denoted by $\widehat{G}$, is called the ***profinite completion*** of G. There is a map $G \to \widehat{G}$, namely, $g \mapsto (gN)$, and it is an injection if and only if G is ***residually finite***; that is, $\bigcap_{N \in \mathcal{N}} N = \{1\}$. It is known, for example, that every free group is residually finite.

There are some lovely results obtained making use of profinite completions. If r is a positive integer, a group G is said to have ***rank*** r if every subgroup of G can be generated by r or fewer elements. If G is a residually finite p-group (every element in G has order a power of p) of rank r, then G is isomorphic to a subgroup of $\text{GL}(n, \mathbb{Z}_p)$ for some n (not every residually finite group admits such a linear imbedding). See Dixon–du Sautoy–Mann–Segal, *Analytic Pro-p Groups*, page 98. ◂

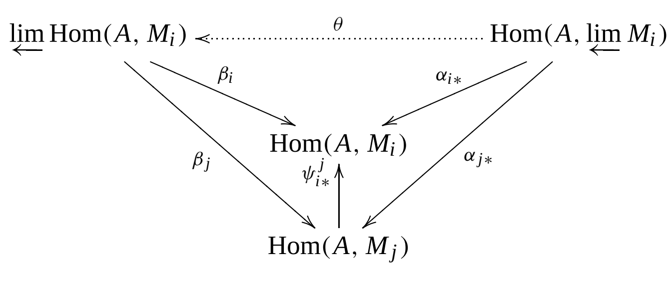
The next result generalizes Theorem 7.32.

Proposition 7.92. *If $\{M_i, \psi_i^j\}$ is an inverse system, then*

$$\text{Hom}(A, \varprojlim M_i) \cong \varprojlim \text{Hom}(A, M_i)$$

for every module A.

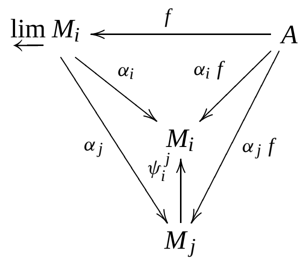
Proof. This statement follows from inverse limit being the solution of a universal mapping problem. In more detail, consider the diagram



where the β_i are the maps given in the definition of inverse limit.

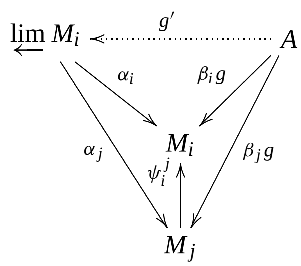
To see that $\theta\colon \text{Hom}(A, \varprojlim M_i) \to \varprojlim \text{Hom}(A, M_i)$ is injective, suppose that $f : A \to \varprojlim M_i$ and $\theta(f) = 0$. Then $0 = \beta_i \theta f = \alpha_i f$ for all i, and so the following diagram

commutes:



But the zero map in place of f also makes the diagram commute, and so the uniqueness of such a map gives $f = 0$; that is, θ is injective.

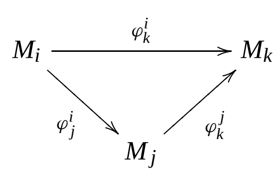
To see that θ is surjective, take $g \in \varprojlim \operatorname{Hom}(A, M_i)$. For each i, there is a map $\beta_i g \colon A \to M_i$ with $\psi_i^j \beta_i g = \beta_j g$.



The definition of $\varprojlim M_i$ provides a map $g' \colon A \to \varprojlim M_i$ with $\alpha_i g' = \beta_i g$ for all i. It follows that $g = \theta(g')$; that is, θ is surjective. •

We now consider the dual construction.

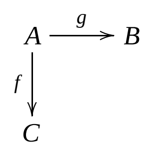
Definition. Let I be a partially ordered set. A ***direct system of*** R***-modules*** over I is an ordered pair $\{M_i, \varphi_j^i\}$ consisting of an indexed family of modules $\{M_i : i \in I\}$ together with a family of morphisms $\{\varphi_j^i \colon M_i \to M_j\}$ for $i \preceq j$, such that $\varphi_i^i = 1_{M_i}$ for all i and such that the following diagram commutes whenever $i \preceq j \preceq k$:



If we regard I as the category $\mathbf{PO}(I)$ whose only morphisms are κ_j^i when $i \preceq j$, and if we define $F(i) = M_i$ and $F(\kappa_j^i) = \varphi_j^i$, then it is easy to see that $\{M_i, \varphi_j^i\}$ is a direct system if and only if $F \colon \mathbf{PO}(I) \to {}_R\mathbf{Mod}$ is a (covariant) functor. Thus, we can consider direct systems involving objects and morphisms in any category $\mathcal{C}$ as being a (covariant) functor

$F : \mathbf{PO}(I) \to \mathcal{C}$. For example, it makes sense to consider direct systems of commutative rings.

Example 7.93.
(i) If $I = \{1, 2, 3\}$ is the partially ordered set in which $1 \preceq 2$ and $1 \preceq 3$, then a direct system over I is a diagram of the form



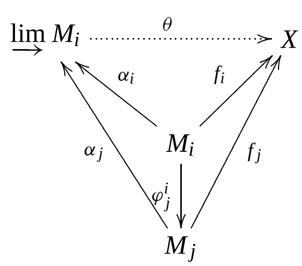
(ii) If $\mathcal{I}$ is a family of submodules of a module A, then it can be partially ordered under inclusion; that is, $M \preceq M'$ in case $M \subseteq M'$. For $M \preceq M'$, the inclusion map $M \to M'$ is defined, and it is easy to see that the family of all $M \in \mathcal{I}$ with inclusion maps is a direct system.

(iii) If I is equipped with the discrete partial order, then a direct system over I is just a family of modules indexed by I. ◂

Definition. Let I be a partially ordered set, and let $\{M_i, \varphi^i_j\}$ be a direct system of R-modules over I. The ***direct limit*** (also called ***inductive limit*** or ***colimit***) is an R-module $\varinjlim M_i$ and a family of R-maps $\{\alpha_i : M_i \to \varinjlim M_i : i \in I\}$, such that

(i) $\alpha_j \varphi^i_j = \alpha_i$ whenever $i \preceq j$;

(ii) for every module X having maps $f_i : M_i \to X$ satisfying $f_j \varphi^i_j = f_i$ for all $i \preceq j$, there exists a unique map $\theta : \varinjlim M_i \to X$ making the following diagram commute:



The notation $\varinjlim M_i$ for a direct limit is deficient in that it does not display the maps of the corresponding direct system (and $\varinjlim M_i$ does depend on them). However, this is standard practice.

As with any object defined as a solution to a universal mapping problem, the direct limit of a direct system is unique (to isomorphism) if it exists.

Proposition 7.94. *The direct limit of any direct system $\{M_i, \varphi^i_j\}$ of R-modules over a partially ordered index set I exists.*

Proof. For each $i \in I$, let λ_i be the injection of M_i into the sum $\sum_i M_i$. Define

$$D = \Big(\sum_i M_i\Big)/S,$$

where S is the submodule of $\sum M_i$ generated by all elements $\lambda_j \varphi^i_j m_i - \lambda_i m_i$ with $m_i \in M_i$ and $i \preceq j$. Now define $\alpha_i : M_i \to D$ by

$$\alpha_i : m_i \mapsto \lambda_i(m_i) + S.$$

It is routine to check that $D \cong \varinjlim M_i$. •

Thus, each element of $\varinjlim M_i$ has a representative of the form $\sum \lambda_i m_i + S$.

The argument in Proposition 7.94 can be modified to prove that direct limits in other categories exist; for example, direct limits of commutative rings, of groups, or of topological spaces exist.

The reader should supply verifications of the following assertions, in which we describe the direct limit of some of the direct systems in Example 7.93.

Example 7.95.

(i) If I is the partially ordered set $\{1, 2, 3\}$ with $1 \preceq 2$ and $1 \preceq 3$, then a direct system is a diagram

$$\begin{array}{ccc} A & \xrightarrow{g} & B \\ {\scriptstyle f}\downarrow & & \\ C & & \end{array}$$

and the direct limit is the pushout.

(ii) If I is a discrete index set, then the direct system is just the indexed family $\{M_i : i \in I\}$, and the direct limit is the sum: $\varinjlim M_i \cong \sum_i M_i$, for the submodule S in the construction of $\varinjlim M_i$ is $\{0\}$. Alternatively, this is just the diagrammatic definition of a coproduct. ◂

The next result generalizes Theorem 7.33.

Proposition 7.96. *If $\{M_i, \varphi^i_j\}$ is a direct system, then*

$$\operatorname{Hom}(\varinjlim M_i, B) \cong \varprojlim \operatorname{Hom}(M_i, B)$$

for every module B.

Proof. This statement follows from direct limit being the solution of a universal mapping problem. The proof is dual to that of Proposition 7.92 and it is left to the reader. •

There is a special kind of partially ordered index set that is useful for direct limits.

Definition. A ***directed set*** is a partially ordered set I such that, for every $i, j \in I$, there is $k \in I$ with $i \preceq k$ and $j \preceq k$.

Example 7.97.

(i) Let $\mathcal{I}$ be a simply ordered family of submodules of a module A; that is, if $M, M' \in \mathcal{I}$, then either $M \subseteq M'$ or $M' \subseteq M$. As in Example 7.93(ii), $\mathcal{I}$ is a partially ordered set; here, $\mathcal{I}$ is a directed set.

(ii) If I is the partially ordered set $\{1, 2, 3\}$ with $1 \preceq 2$ and $1 \preceq 3$, then I is *not* a directed set.

(iii) If $\{M_i : i \in I\}$ is some family of modules, and if I is a discrete partially ordered index set, then I is not directed. However, if we consider the family $\mathcal{F}$ of all ***finite partial sums***

$$M_{i_1} \oplus \cdots \oplus M_{i_n},$$

then $\mathcal{F}$ is a directed set under inclusion.

(iv) If A is a module, then the family $\text{Fin}(A)$ of all the finitely generated submodules of A is partially ordered by inclusion, as in Example 7.93(ii), and it is a directed set.

(v) If R is a domain and $Q = \text{Frac}(R)$, then the family of all cyclic R-submodules of Q of the form $\langle 1/r \rangle$, where $r \in R$ and $r \neq 0$, is a partially ordered set, as in Example 7.93(ii); here, it is a directed set under inclusion, for given $\langle 1/r \rangle$ and $\langle 1/s \rangle$, then each is contained in $\langle 1/rs \rangle$.

(vi) Let $\mathcal{U}$ be the family of all the open intervals in $\mathbb{R}$ containing 0. Partially order $\mathcal{U}$ by reverse inclusion:

$$U \preceq V \quad \text{if} \quad V \subseteq U.$$

Notice that $\mathcal{U}$ is directed: Given $U, V \in \mathcal{U}$, then $U \cap V \in \mathcal{U}$ and $U \preceq U \cap V$ and $V \preceq U \cap V$.

For each $U \in \mathcal{U}$, define

$$\mathcal{F}(U) = \{ f : U \to \mathbb{R} : f \text{ is continuous}\},$$

and, if $U \preceq V$, that is, $V \subseteq U$, define $\rho^U_V : \mathcal{F}(U) \to \mathcal{F}(V)$ to be the restriction map $f \mapsto f|V$. Then $\{\mathcal{F}(U), \rho^U_V\}$ is a direct system. ◀

There are two reasons to consider direct systems with directed index sets. The first is that a simpler description of the elements in the direct limit can be given; the second is that $\varinjlim$ preserves short exact sequences.

Proposition 7.98. *Let $\{M_i, \varphi^i_j\}$ be a direct system of left R-modules over a directed index set I, and let $\lambda_i : M_i \to \sum M_i$ be the ith injection, so that $\varinjlim M_i = (\sum M_i)/S$, where*

$$S = \big\langle \lambda_j \varphi^i_j m_i - \lambda_i m_i : m_i \in M_i \text{ and } i \preceq j \big\rangle.$$

(i) *Each element of $\varinjlim M_i$ has a representative of the form $\lambda_i m_i + S$ (instead of $\sum_i \lambda_i m_i + S$).*

(ii) $\lambda_i m_i + S = 0$ *if and only if* $\varphi_t^i(m_i) = 0$ *for some* $t \succeq i$.

Proof. (i) As in the proof Proposition 7.94, the existence of direct limits, $\varinjlim M_i = (\sum M_i)/S$, and so a typical element $x \in \varinjlim M_i$ has the form $x = \sum \lambda_i m_i + S$. Since I is directed, there is an index j with $j \succeq i$ for all i occurring in the sum for x. For each such i, define $b^i = \varphi_j^i m_i \in M_j$, so that the element b, defined by $b = \sum_i b^i$ lies in M_j. It follows that

$$\begin{aligned}\sum \lambda_i m_i - \lambda_j b &= \sum(\lambda_i m_i - \lambda_j b^i)\\ &= \sum(\lambda_i m_i - \lambda_j \varphi_j^i m_i) \in S.\end{aligned}$$

Therefore, $x = \sum \lambda_i m_i + S = \lambda_j b + S$, as desired.

(ii) If $\varphi_t^i m_i = 0$ for some $t \succeq i$, then

$$\lambda_i m_i + S = \lambda_i m_i + (\lambda_t \varphi_t^i m_i - \lambda_i m_i) + S = S.$$

Conversely, if $\lambda_i m_i + S = 0$, then $\lambda_i m_i \in S$, and there is an expression

$$\lambda_i m_i = \sum_j a_j(\lambda_k \varphi_k^j m_j - \lambda_j m_j) \in S,$$

where $a_j \in R$. We are going to normalize this expression; first, we introduce the following notation for relators: If $j \preceq k$, define

$$r(j, k, m_j) = \lambda_k \varphi_k^j m_j - \lambda_j m_j.$$

Since $a_j r(j, k, m_j) = r(j, k, a_j m_j)$, we may assume that the notation has been adjusted so that

$$\lambda_i m_i = \sum_j r(j, k, m_j).$$

As I is directed, we may choose an index $t \in I$ larger than any of the indices i, j, k occurring in the last equation. Now

$$\begin{aligned}\lambda_t \varphi_t^i m_i &= (\lambda_t \varphi_t^i m_i - \lambda_i m_i) + \lambda_i m_i\\ &= r(i, t, m_i) + \lambda_i m_i\\ &= r(i, t, m_i) + \sum_j r(j, k, m_j).\end{aligned}$$

Next,

$$\begin{aligned}r(j, k, m_j) &= \lambda_k \varphi_k^j m_j - \lambda_j m_j\\ &= (\lambda_t \varphi_t^j m_j - \lambda_j m_j) + \Big[\lambda_t \varphi_t^k(-\varphi_k^j m_j) - \lambda_k(-\varphi_k^j m_j)\Big]\\ &= r(j, t, m_j) + r(k, t, -\varphi_k^j m_j),\end{aligned}$$

because $\varphi_t^k \varphi_k^i = \varphi_t^i$, by definition of direct system. Hence,

$$\lambda_t \varphi_t^i m_i = \sum_\ell r(\ell, t, x_\ell),$$

where $x_\ell \in M_\ell$. But it is easily checked, for $\ell \preceq t$, that

$$r(\ell, t, m_\ell) + r(\ell, t, m'_\ell) = r(\ell, t, m_\ell + m'_\ell).$$

Therefore, we may amalgamate all relators with the same smaller index ℓ and write

$$\begin{aligned}\lambda_t \varphi_t^i m_i &= \sum_\ell r(\ell, t, x_\ell) \\ &= \sum_\ell \lambda_t \varphi_t^\ell x_\ell - \lambda_\ell x_\ell \\ &= \lambda_t \big(\sum_\ell \varphi_t^\ell x_\ell\big) - \sum_\ell \lambda_\ell x_\ell,\end{aligned}$$

where $x_\ell \in M_\ell$ and all the indices ℓ are distinct. The unique expression of an element in a direct sum allows us to conclude, if $\ell \neq t$, that $\lambda_\ell x_\ell = 0$; it follows that $x_\ell = 0$, for λ_ℓ is an injection. The right side simplifies to $\lambda_t \varphi_t^t m_t - \lambda_t m_t = 0$, because φ_t^t is the identity. Thus, the right side is 0 and $\lambda_t \varphi_t^i m_i = 0$. Since λ_t is an injection, we have $\varphi_t^i m_i = 0$, as desired. •

Our original construction of $\varinjlim M_i$ involved a quotient of $\sum M_i$; that is, $\varinjlim M_i$ is a quotient of a coproduct. In the category **Sets**, coproduct is disjoint union $\bigsqcup_i M_i$. We may regard a "quotient" of a set X as the family of equivalence classes of some equivalence relation on X. This categorical analogy suggests that we might be able to give a second construction of $\varinjlim M_i$ using an equivalence relation on $\bigsqcup_i M_i$. When the index set is directed, this can actually be done (see Exercise 7.65 on page 517).

Example 7.99.
(i) Let $\mathcal{I}$ be a simply ordered family of submodules of a module A; that is, if $M, M' \in \mathcal{I}$, then either $M \subseteq M'$ or $M' \subseteq M$. Then $\mathcal{I}$ is a directed set, and $\varinjlim M_i \cong \bigcup_i M_i$.

(ii) If $\{M_i : i \in I\}$ is some family of modules, then $\mathcal{F}$, all ***finite partial sums***, is a directed set under inclusion, and $\varinjlim M_i \cong \sum_i M_i$.

(iii) If A is a module, then the family $\mathrm{Fin}(A)$ of all the finitely generated submodules of A is a directed set and $\varinjlim M_i \cong A$.

(iv) If R is a domain and $Q = \mathrm{Frac}(R)$, then the family of all cyclic R-submodules of Q of the form $\langle 1/r \rangle$, where $r \in R$ and $r \neq 0$, forms a directed set under inclusion, and $\varinjlim M_i \cong Q$; that is, Q is a direct limit of cyclic modules. ◂

Definition. Let $\{A_i, \alpha^i_j\}$ and $\{B_i, \beta^i_j\}$ be direct systems over the same index set I. A ***transformation*** $r\colon \{A_i, \alpha^i_j\} \to \{B_i, \beta^i_j\}$ is an indexed family of homomorphisms

$$r = \{r_i\colon A_i \to B_i\}$$

that makes the following diagram commute for all $i \preceq j$:

$$\begin{array}{ccc} A_i & \xrightarrow{r_i} & B_i \\ {\scriptstyle \alpha^i_j}\downarrow & & \downarrow{\scriptstyle \beta^i_j} \\ A_j & \xrightarrow[r_j]{} & B_j \end{array}$$

A transformation $r\colon \{A_i, \alpha^i_j\} \to \{B_i, \beta^i_j\}$ determines a homomorphism

$$\vec{r}\colon \varinjlim A_i \to \varinjlim B_i$$

by

$$\vec{r}\colon \sum \lambda_i a_i + S \mapsto \sum \mu_i r_i a_i + T,$$

where $S \subseteq \sum A_i$ and $T \subseteq \sum B_i$ are the relation submodules in the construction of $\varinjlim A_i$ and $\varinjlim B_i$, respectively, and λ_i and μ_i are the injections of A_i and B_i into the direct sums. The reader should check that r being a transformation of direct systems implies that $\vec{r}$ is independent of the choice of coset representative, and hence it is a well-defined function.

Proposition 7.100. *Let I be a directed set, and let $\{A_i, \alpha^i_j\}$, $\{B_i, \beta^i_j\}$, and $\{C_i, \gamma^i_j\}$ be direct systems over I. If $r\colon \{A_i, \alpha^i_j\} \to \{B_i, \beta^i_j\}$ and $s\colon \{B_i, \beta^i_j\} \to \{C_i, \gamma^i_j\}$ are transformations, and if*

$$0 \to A_i \xrightarrow{r_i} B_i \xrightarrow{s_i} C_i \to 0$$

is exact for each $i \in I$, then there is an exact sequence

$$0 \to \varinjlim A_i \xrightarrow{\vec{r}} \varinjlim B_i \xrightarrow{\vec{s}} \varinjlim C_i \to 0.$$

Remark. The hypothesis that I be directed enters the proof only in showing that $\vec{r}$ is an injection. ◀

Proof. We prove only that $\vec{r}$ is an injection, for the proof of exactness of the rest is routine. Suppose that $\vec{r}(x) = 0$, where $x \in \varinjlim A_i$. Since I is directed, Proposition 7.98(i) allows us to write $x = \lambda_i a_i + S$ (where $S \subseteq \sum A_i$ is the relation submodule and λ_i is the injection of A_i into the direct sum). By definition, $\vec{r}(x + S) = \mu_i r_i a_i + T$ (where $T \subseteq \sum B_i$ is the relation submodule and μ_i is the injection of B_i into the direct sum). Now Proposition 7.98(ii) shows that $\mu_i r_i a_i + T = 0$ in $\varinjlim B_i$ implies that there is an index $k \succeq i$ with $\beta^i_k r_i a_i = 0$. Since r is a transformation of direct systems, we have

$$0 = \beta^i_k r_i a_i = r_k \alpha^i_k a_i.$$

Finally, since r_k is an injection, we have $\alpha^i_k a_i = 0$ and, hence, that $x = \lambda_i a_i + S = 0$. Therefore, $\vec{r}$ is an injection. •

Example 7.101.
Let $\mathcal{U}$ be the family of all the open intervals in $\mathbb{R}$ containing 0, partially ordered by reverse inclusion, and let $\{B(U), \beta^U_V\}$ be the direct system of Example 7.97(vi), where

$$B(U) = \{ f : U \to \mathbb{R} : f \text{ is continuous}\}$$

and $\beta^U_V : f \mapsto f|V$.

We now present two more direct systems over $\mathcal{U}$. Define

$$A(U) = \{\text{constant functions } f : U \to \mathbb{Z}\}$$

and

$$C(U) = \{\text{continuous } f : U \to \mathbb{R} - \{0\}\},$$

the abelian group under pointwise multiplication. Then $\{A(U), \alpha^U_V\}$ and $\{C(U), \gamma^U_V\}$ are direct systems, where the α and γ are restriction maps.

Define transformations $s : \{B(U), \beta^U_V\} \to \{C(U), \gamma^U_V\}$ by setting $s(U) : B(U) \to C(U)$ to be the map $f \mapsto e^{2\pi i f}$, and define $r : \{A(U), \alpha^U_V\} \to \{B(U), \beta^U_V\}$ by setting $r(U) : A(U) \to B(U)$ to be the inclusion map. It is easy to see that

$$0 \to A(U) \overset{r_U}{\to} B(U) \overset{s_U}{\to} C(U) \to 0$$

is exact for all U, and so Proposition 7.100 gives exactness of

$$0 \to \varinjlim A(U) \to \varinjlim B(U) \to \varinjlim C(U) \to 0.$$

It is easy to check that $\varinjlim A(U) \cong \mathbb{Z}$, and so $\varinjlim B(U) \neq 0$. ◂

There is a way to compare two functors.

Definition. Let $F : \mathcal{C} \to \mathcal{D}$ and $G : \mathcal{C} \to \mathcal{D}$ be covariant functors. A ***natural transformation*** is a family of morphisms $\tau = \{\tau_C : FC \to GC\}$, one for each object C in $\mathcal{C}$, so that the following diagram commutes for all $f : C \to C'$ in $\mathcal{C}$:

$$\begin{array}{ccc} FC & \xrightarrow{Ff} & FC' \\ {\scriptstyle \tau_C}\downarrow & & \downarrow{\scriptstyle \tau_{C'}} \\ GC & \xrightarrow{Gf} & GC' \end{array}$$

If each τ_C is an equivalence, then τ is called a ***natural equivalence*** and F and G are called ***naturally equivalent***.

There is a similar definition of natural transformation between contravariant functors.

The next proposition shows that the isomorphisms $\varphi_M : \mathrm{Hom}_R(R, M) \to M$ in Exercise 7.5 on page 440 constitute a natural transformation.

Proposition 7.102. *If R is a commutative ring, then* $\mathrm{Hom}_R(R, M)$ *is an R-module, and the R-isomorphisms*

$$\varphi_M \colon \mathrm{Hom}_R(R, M) \to M,$$

given by $\varphi_M(f) = f(1)$, comprise a natural equivalence $\varphi\colon \mathrm{Hom}_R(R, \quad) \to 1_R$, the identity functor on ${}_R\mathbf{Mod}$.

Remark. Proposition 8.85 generalizes this proposition to modules over noncommutative rings. ◂

Proof. It is easy to check that φ_M is an additive function. To see that φ_M is an R-homomorphism, note that

$$\varphi_M(rf) = (rf)(1) = f(1r) = f(r) = r[f(1)] = r\varphi_M(f),$$

because f is an R-map. Consider the function $M \to \mathrm{Hom}_R(R, M)$ defined as follows: If $m \in M$, then $f_m \colon R \to M$ is given by $f_m(r) = rm$; it is easy to see that f_m is an R-homomorphism, and that $m \mapsto f_m$ is the inverse of φ_M.

To see that the isomorphisms φ_M constitute a natural equivalence, it suffices to show, for any module homomorphism $h \colon M \to N$, that the following diagram commutes:

$$\begin{array}{ccc} \mathrm{Hom}_R(R, M) & \xrightarrow{h_*} & \mathrm{Hom}_R(R, N) \\ {\scriptstyle \varphi_M}\downarrow & & \downarrow{\scriptstyle \varphi_N} \\ M & \xrightarrow{h} & N, \end{array}$$

where $h_* \colon f \mapsto hf$. Let $f \colon R \to M$. Going clockwise, $f \mapsto hf \mapsto hf(1)$, while going counterclockwise, $f \mapsto f(1) \mapsto h(f(1))$. •

An analysis of the proof of Proposition 7.92 shows that it can be generalized by replacing $\mathrm{Hom}(A, \quad)$ by any (covariant) left exact functor $F \colon {}_R\mathbf{Mod} \to \mathbf{Ab}$ that preserves products. However, this added generality is only illusory, for it is a theorem of C. E. Watts, given such a functor F, that there exists a module A with F naturally equivalent to $\mathrm{Hom}_R(A, \quad)$; that is, these representable functors are characterized. Another theorem of Watts characterizes contravariant functors: If $G \colon {}_R\mathbf{Mod} \to \mathbf{Ab}$ is a contravariant left exact functor that converts sums to products, then there exists a module B with G naturally equivalent to $\mathrm{Hom}_R(\quad, B)$. Proofs of Watts's theorems can be found in Rotman, *An Introduction to Homological Algebra*, pages 77–79.

Example 7.103.

(i) In Proposition 7.100, we introduced transformations from one direct system over a partially ordered index set I to another. If we recall that a direct system of R-modules over I can be regarded as a functor $\mathbf{PO}(I) \to_R \mathbf{Mod}$, then the reader can see that these transformations are natural transformations.

If we regard inverse systems over a partially ordered index set as contravariant functors, then we can also define transformations between them (as natural transformations).

(ii) Choose a point p once for all, and let $P = \{p\}$; we claim that $\text{Hom}(P, \quad)\colon \textbf{Sets} \to \textbf{Sets}$ is naturally equivalent to the identity functor on **Sets**. If X is a set, define

$$\tau_X\colon \text{Hom}(P, X) \to X \text{ by } f \mapsto f(p).$$

Each τ_X is a bijection, as is easily seen, and we now show that τ is a natural transformation. Let X and Y be sets, and let $h\colon X \to Y$; we must show that the following diagram commutes:

$$\begin{array}{ccc} \text{Hom}(P, X) & \xrightarrow{h_*} & \text{Hom}(P, Y) \\ {\scriptstyle \tau_X}\downarrow & & \downarrow{\scriptstyle \tau_Y} \\ X & \xrightarrow{h} & Y, \end{array}$$

where $h_*\colon f \mapsto hf$. Going clockwise, $f \mapsto hf \mapsto hf(p)$, while going counterclockwise, $f \mapsto f(p) \mapsto h(f(p))$.

(iii) If k is a field and V is a vector space over k, then its dual space V^* is the vector space $\text{Hom}_k(V, k)$ of all linear functionals on V. The evaluation map $e_v\colon f \mapsto f(v)$ is a linear functional on V^*; that is, $e_v \in (V^*)^* = V^{**}$. Define $\tau_V\colon V \to V^{**}$ by

$$\tau_V\colon v \mapsto e_v.$$

The reader may check that τ is a natural transformation from the identity functor on ${}_k\textbf{Mod}$ to the double dual functor. The restriction of τ to the subcategory of all finite-dimensional vector spaces is a natural equivalence. ◂

There is a lovely part of ring theory developing these ideas. The first question is when a category $\mathcal{C}$ is "isomorphic" to a category ${}_R\textbf{Mod}$ of modules; we have to be a bit fussy about what isomorphism means here; it is a bit weaker than having functors $F\colon \mathcal{C} \to {}_R\textbf{Mod}$ and $G\colon {}_R\textbf{Mod} \to \mathcal{C}$ with both composites equal to identity functors.

Definition. A functor $F\colon \mathcal{C} \to \mathcal{D}$ is an ***equivalence*** if there is a functor $G\colon \mathcal{D} \to \mathcal{C}$ such that the composites GF and FG are naturally equivalent to the identity functors $1_{\mathcal{C}}$ and $1_{\mathcal{D}}$, respectively.

Morita theory proves that if R and S are commutative rings, then equivalence of their module categories implies $R \cong S$. We will say a few words about Morita theory in Chapter 9, once we introduce modules over noncommutative rings, but the reader should really read accounts of Morita theory in Jacobson, *Basic Algebra* II or in Lam, *Lectures on Modules and Rings*.

Definition. Given functors $F\colon \mathcal{C} \to \mathcal{D}$ and $G\colon \mathcal{D} \to \mathcal{C}$, then the ordered pair (F, G) is called an ***adjoint pair*** if, for each pair of objects $C \in \mathcal{C}$ and $D \in \mathcal{D}$, there are bijections

$$\tau_{C,D}\colon \text{Hom}_{\mathcal{D}}(FC, D) \to \text{Hom}_{\mathcal{C}}(C, GD)$$

that are natural transformations in C and in D.

In more detail, the following two diagrams commute: For every $f: C' \to C$ in $\mathcal{C}$ and $g: D \to D'$ in $\mathcal{D}$,

$$\begin{array}{ccc}
\operatorname{Hom}_{\mathcal{D}}(FC, D) & \xrightarrow{(Ff)^*} & \operatorname{Hom}_{\mathcal{D}}(FC', D) \\
\downarrow{\scriptstyle \tau_{C,D}} & & \downarrow{\scriptstyle \tau_{C',D}} \\
\operatorname{Hom}_{\mathcal{C}}(C, GD) & \xrightarrow{f^*} & \operatorname{Hom}_{\mathcal{C}}(C', GD);
\end{array}$$

$$\begin{array}{ccc}
\operatorname{Hom}_{\mathcal{D}}(FC, D) & \xrightarrow{g_*} & \operatorname{Hom}_{\mathcal{D}}(FC, D') \\
\downarrow{\scriptstyle \tau_{C,D}} & & \downarrow{\scriptstyle \tau_{C,D'}} \\
\operatorname{Hom}_{\mathcal{C}}(C, GD) & \xrightarrow{(Gg)_*} & \operatorname{Hom}_{\mathcal{C}}(C, GD').
\end{array}$$

Here is the etymology of "adjoint." Let $F = \quad \otimes_R B: \mathbf{Mod}_R \to \mathbf{Mod}_S$, and let $G = \operatorname{Hom}_S(B, \quad): \mathbf{Mod}_S \to \mathbf{Mod}_R$. The isomorphism in Theorem 8.99 is

$$\tau: \operatorname{Hom}_S(F(A), C) \to \operatorname{Hom}_R(A, G(C)).$$

If we pretend that $\operatorname{Hom}(\ ,\ )$ is an inner product, then this reminds us of the definition of adjoint pairs in linear algebra: If $T: V \to W$ is a linear transformation of vector spaces equipped with inner products, then its adjoint is the linear transformation $T^*: W \to V$ such that

$$(Tv, w) = (v, T^*w)$$

for all $v \in V$ and $w \in W$.

Example 7.104.

(i) Let U: **Groups** $\to$ **Sets** be the ***underlying functor***, which assigns to each group G its underlying set and views each homomorphism as a mere function, and let F: **Sets** $\to$ **Groups** be the ***free functor***, which assigns to each set X the free group FX having basis X. That FX is free with basis X says, for every group H, that every function $\varphi: X \to H$ corresponds to a unique homomorphism $\widetilde{\varphi}: FX \to H$. It follows that if $\varphi: X \to Y$ is any function, then $\widetilde{\varphi}: FX \to FY$; this is how F is defined on morphisms: $F\varphi = \widetilde{\varphi}$. The reader should realize that the function $f \mapsto f|X$ is a bijection (whose inverse is $\varphi \mapsto \widetilde{\varphi}$)

$$\tau_{X,H}: \operatorname{Hom}_{\mathbf{Groups}}(FX, H) \to \operatorname{Hom}_{\mathbf{Sets}}(X, UH).$$

Indeed, $\tau_{X,H}$ is a natural bijection, showing that (F, U) is an adjoint pair of functors.

This example can be generalized by replacing **Groups** by other categories having free objects; for example, ${}_R\mathbf{Mod}$ for any ring R.

(ii) Adjointness is a property of an *ordered pair* of functors. In (i), we saw that (F, U) is an adjoint pair, where F is a free functor and U is the underlying functor. Were (U, F) an adjoint pair, then there would be a natural bijection $\operatorname{Hom}_{\mathbf{Sets}}(UH, Y) \cong \operatorname{Hom}_{\mathbf{Groups}}(H, FY)$, where H is a group and Y is a set. This is false in general; if H is a finite group with more

than one element and Y is a set with more than one element, then $\text{Hom}_{\textbf{Sets}}(UH, Y)$ has more than one element, but $\text{Hom}_{\textbf{Groups}}(H, FY)$ has only one element. Therefore, (U, F) is not an adjoint pair.

(iii) In the next chapter, we shall see (Theorem 8.99) that for every covariant Hom functor $G = \text{Hom}_R(A, \ )$, there exists a functor F such that (F, G) is an adjoint pair ($F = A \otimes_R$ is called *tensor product*). ◂

For many more examples of adjoint pairs of functors, see Mac Lane, *Categories for the Working Mathematician*, Chapter 4, especially pages 85–86.

Let (F, G) be an adjoint pair of functors, where $F : \mathcal{C} \to \mathcal{D}$ and $G : \mathcal{D} \to \mathcal{C}$. If $C \in \text{obj}(\mathcal{C})$, then setting $D = FC$ gives a bijection $\tau : \ \text{Hom}_{\mathcal{D}}(FC, FC) \to \text{Hom}_{\mathcal{C}}(C, GFC)$, so that η_C, defined by

$$\eta_C = \tau(1_{FC}),$$

is a morphism $C \to GFC$. Exercise 7.75 on page 518 shows that $\eta : 1_{\mathcal{C}} \to GF$ is a natural transformation; it is called the ***unit*** of the adjoint pair.

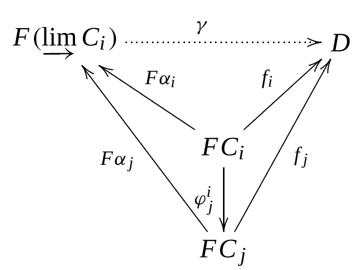
Theorem 7.105. *Let (F, G) be an adjoint pair of functors, where $F : \mathcal{C} \to \mathcal{D}$ and $G : \mathcal{D} \to \mathcal{C}$. Then F preserves all direct limits and G preserves all inverse limits.*

Remark.

(i) There is no restriction on the index sets of the limits; in particular, they need not be directed.

(ii) A more precise statement is that if $\varinjlim C_i$ exists in $\mathcal{C}$, then $\varinjlim FC_i$ exists in $\mathcal{D}$, and $\varinjlim FC_i \cong F(\varinjlim C_i)$. ◂

Proof. Let I be a partially ordered set, and let $\{C_i, \varphi^i_j\}$ be a direct system in $\mathcal{C}$ over I. It is easy to see that $\{FC_i, F\varphi^i_j\}$ is a direct system in $\mathcal{D}$ over I. Consider the following diagram in $\mathcal{D}$:



where $\alpha_i : C_i \to \varinjlim C_i$ are the maps in the definition of direct limit. We must show that there exists a unique morphism $\gamma : F(\varinjlim C_i) \to D$ making the diagram commute. The idea is to apply G to this diagram, and to use the unit $\eta : 1_{\mathcal{C}} \to GF$ to replace $GF(\varinjlim C_i)$

and GFC_i by $\varinjlim C_i$ and C_i, respectively. In more detail, there are morphisms η and η_i, by Exercise 7.75 on page 518, making the following diagram commute:

$$\begin{array}{ccc} \varinjlim C_i & \xrightarrow{\eta} & GF(\varinjlim C_i) \\ {\scriptstyle \alpha_i}\uparrow & & \uparrow{\scriptstyle GF\alpha_i} \\ C_i & \xrightarrow{\eta_i} & GFC_i \end{array}$$

Combining this with G applied to the original diagram gives commutativity of

$$\begin{array}{ccc} \varinjlim C_i & \overset{\beta}{\dashrightarrow} & GD \\ & {\scriptstyle \alpha_i}\nwarrow \quad C_i \quad \nearrow{\scriptstyle (Gf_i)\eta_i} & \\ {\scriptstyle \alpha_j}\nwarrow & \downarrow{\scriptstyle \varphi^i_j} & \nearrow{\scriptstyle (Gf_j)\eta_j} \\ & C_j & \end{array}$$

By definition of direct limit, there exists a unique $\beta\colon \varinjlim C_i \to GD$ making the diagram commute; that is, $\beta \in \operatorname{Hom}_{\mathcal{C}}(\varinjlim C_i, GD)$. Since (F, G) is an adjoint pair, there exists a natural bijection

$$\tau\colon \operatorname{Hom}_{\mathcal{D}}(F(\varinjlim C_i), D) \to \operatorname{Hom}_{\mathcal{C}}(\varinjlim C_i, GD).$$

Define

$$\gamma = \tau^{-1}(\beta) \in \operatorname{Hom}_{\mathcal{D}}(F(\varinjlim C_i), D).$$

We claim that $\gamma\colon F(\varinjlim C_i) \to D$ makes the first diagram commute. The first commutative square in the definition of adjointness gives commutativity of

$$\begin{array}{ccc} \operatorname{Hom}_{\mathcal{C}}(\varinjlim C_i, GD) & \xrightarrow{\alpha_i^*} & \operatorname{Hom}_{\mathcal{C}}(C_i, GD) \\ {\scriptstyle \tau^{-1}}\downarrow & & \downarrow{\scriptstyle \tau^{-1}} \\ \operatorname{Hom}_{\mathcal{D}}(F(\varinjlim C_i), D) & \xrightarrow{(F\alpha_i)^*} & \operatorname{Hom}_{\mathcal{D}}(FC_i, D). \end{array}$$

Hence, $\tau^{-1}\alpha_i^* = (F\alpha_i)^*\tau^{-1}$. Evaluating both functions on β, we have

$$(F\alpha_i)^*\tau^{-1}(\beta) = (F\alpha_i)^*\gamma = \gamma F\alpha_i.$$

On the other hand, since $\beta\alpha_i = (Gf_i)\eta_i$, we have

$$\tau^{-1}\alpha_i^*(\beta) = \tau^{-1}(\beta\alpha_i) = \tau^{-1}((Gf_i)\eta_i).$$

Therefore,

$$\gamma F\alpha_i = \tau^{-1}((Gf_i)\eta_i).$$

The second commutative square in the definition of adjointness gives commutativity of

$$\begin{array}{ccc} \mathrm{Hom}_{\mathcal{D}}(FC_i, FC_i) & \xrightarrow{(f_i)_*} & \mathrm{Hom}_{\mathcal{D}}(FC_i, D) \\ \tau \downarrow & & \downarrow \tau \\ \mathrm{Hom}_{\mathcal{C}}(C_i, GFC_i) & \xrightarrow{(Gf_i)_*} & \mathrm{Hom}_{\mathcal{C}}(C_i, GD); \end{array}$$

that is,

$$\tau(f_i)_* = (Gf_i)_*\tau.$$

Evaluating at 1_{FC_i}, the definition of η_i gives $\tau(f_i)_*(1) = (Gf_i)_*\tau(1)$, and so $\tau f_i = (Gf_i)_*\eta_i$. Therefore,

$$\gamma F\alpha_i = \tau^{-1}((Gf_i)\eta_i) = \tau^{-1}\tau f_i = f_i,$$

so that γ makes the original diagram commute.

We leave the proof of the uniqueness of γ as an exercise for the reader, with the hint to use the uniqueness of β.

The dual proof shows that G preserves inverse limits. •

There is a necessary and sufficient condition, called the ***adjoint functor theorem***, that a functor be part of an adjoint pair; see Mac Lane, *Categories for the Working Mathematician*, page 117.

EXERCISES

7.65 Let $\{M_i, \varphi^i_j\}$ be a direct system of R-modules with index set I, and let $\bigsqcup_i M_i$ be the disjoint union. Define $m_i \sim m_j$ on $\bigsqcup_i M_i$, where $m_i \in M_i$ and $m_j \in M_j$, if there exists an index k with $k \succeq i$ and $k \succeq j$ such that $\varphi^i_k m_i = \varphi^j_k m_j$.

(i) Prove that $\sim$ is an equivalence relation on $\bigsqcup_i M_i$.

(ii) Denote the equivalence class of m_i by $[m_i]$, and let L denote the family of all such equivalence classes. Prove that the following definitions give L the structure of an R-module:

$$r[m_i] = [rm_i] \text{ if } r \in R;$$

$$[m_i] + [m'_j] = [\varphi^i_k m_i + \varphi^j_k m'_j], \text{ where } k \succeq i \text{ and } k \succeq j.$$

(iii) Prove that $L \cong \varinjlim M_i$.

Hint. Use Proposition 7.98.

7.66 Let $\{M_i, \varphi^i_j\}$ be a direct system of R-modules, and let $F\colon {}_R\mathbf{Mod} \to \mathcal{C}$ be a functor to some category $\mathcal{C}$. Prove that $\{FM_i, F\varphi^i_j\}$ is a direct system in $\mathcal{C}$ if F is covariant, while it is an inverse system if F is contravariant.

7.67 Give an example of a direct system of modules, $\{A_i, \alpha^i_j\}$, over some directed index set I, for which $A_i \neq \{0\}$ for all i and $\varinjlim A_i = \{0\}$.

7.68 (i) Let K be a cofinal subset of a directed index set I (that is, for each $i \in I$, there is $k \in K$ with $i \preceq k$), let $\{M_i, \varphi^i_j\}$ be a direct system over I, and let $\{M_i, \varphi^i_j\}$ be the subdirect system whose indices lie in K. Prove that the direct limit over I is isomorphic to the direct limit over K.

(ii) A partially ordered set I has a ***top element*** if there exists $\infty \in I$ with $i \preceq \infty$ for all $i \in I$. If $\{M_i, \varphi^i_j\}$ is a direct system over I, prove that

$$\varinjlim M_i \cong M_\infty.$$

(iii) Show that part (i) may not be true if the index set is not directed.
Hint. Pushout.

7.69 Let $\mathcal{C}$ and $\mathcal{D}$ be categories, and let $\mathcal{F}(\mathcal{C}, \mathcal{D})$ denote the class of all (covariant) functors $\mathcal{C} \to \mathcal{D}$. Prove that $\mathcal{F}(\mathcal{C}, \mathcal{D})$ is a category if we define

$$\mathrm{Hom}(F, G) = \{\text{all natural transformations } F \to G\}.$$

Remark. There is a technical, set-theoretic, problem; why is $\mathrm{Hom}(F, G)$ a set (and not a proper class)? The answer is that it may not be a set; the easiest (but not the only) way to resolve this problem is to assume that the objects in $\mathcal{C}$ and $\mathcal{D}$ form a set; that is, $\mathcal{C}$ and $\mathcal{D}$ are small categories. We allow the reader to do this here.

7.70 A functor $T \colon {}_R\mathbf{Mod} \to \mathbf{Ab}$ is called ***representable*** if there exists an R-module A and a natural equivalence $\tau \colon T \to \mathrm{Hom}_R(A, \ )$. Prove that if $\mathrm{Hom}_R(A, \ )$ and $\mathrm{Hom}_R(B, \ )$ are naturally equivalent, then $A \cong B$. Conclude that if a representable functor T is naturally equivalent to $\mathrm{Hom}_R(A, \ )$, then A is determined, up to isomorphism, by T.

7.71 If ${}_k\mathbf{V}$ is the category of all finite-dimensional vector spaces over a field k, prove that the double dual, $V \mapsto V^{**}$, is naturally equivalent to the identity functor.

7.72 Let $\{E_i, \varphi^i_j\}$ be a direct system of injective R-modules over a directed index set I. Prove that if R is noetherian, then $\varinjlim E_i$ is an injective module.
Hint. Use Proposition 7.69.

7.73 Consider the ideal $I = (x)$ in $k[x]$, where k is a commutative ring. Prove that the completion of the polynomial ring $k[x]$ is $k[[x]]$, the ring of formal power series.

7.74 Let $r \colon \{A_i, \alpha^i_j\} \to \{B_i, \beta^i_j\}$ and $s \colon \{B_i, \beta^i_j\} \to \{C_i, \gamma^i_j\}$ be transformations of inverse systems over an index set I. If

$$0 \to A_i \xrightarrow{r_i} B_i \xrightarrow{s_i} C_i$$

is exact for each $i \in I$, prove that there is an exact sequence

$$0 \to \varprojlim A_i \xrightarrow{\vec{r}} \varprojlim B_i \xrightarrow{\vec{s}} \varprojlim C_i.$$

7.75 Let (F, G) be an adjoint pair of functors, where $F \colon \mathcal{C} \to \mathcal{D}$ and $G \colon \mathcal{D} \to \mathcal{C}$, and let $\tau_{C,D} \colon \mathrm{Hom}(FC, D) \to \mathrm{Hom}(C, GC)$ be the natural bijection.

(i) If $D = FC$, there is a natural bijection

$$\tau_{C,FC} \colon \mathrm{Hom}(FC, FC) \to \mathrm{Hom}(C, GFC)$$

with $\tau(1_{FC}) = \eta_C \in \mathrm{Hom}(C, GFC)$. Prove that $\eta \colon 1_{\mathcal{C}} \to GF$ is a natural transformation.

(ii) If $C = GD$, there is a natural bijection

$$\tau_{GD,D}^{-1} \colon \operatorname{Hom}(GD, GD) \to \operatorname{Hom}(FGD, D)$$

with $\tau^{-1}(1_D) = \varepsilon_D \in \operatorname{Hom}(FGD, D)$. Prove that $\varepsilon \colon FG \to 1_{\mathcal{D}}$ is a natural transformation. (We call ε the ***counit*** of the adjoint pair.)

7.76 If I is a partially ordered set and $\mathcal{C}$ is a category, then a ***presheaf*** over I to $\mathcal{C}$ is a contravariant functor $\mathcal{F} \colon \mathbf{PO}(I) \to \mathcal{C}$.

(i) If I is the family of all open intervals U in $\mathbb{R}$ containing 0, show that $\mathcal{F}$ in Example 7.97(vi) is a presheaf of abelian groups.

(ii) Let X be a topological space, and let I be the partially ordered set whose elements are the open sets in X. Define a sequence of presheaves $\mathcal{F}' \to \mathcal{F} \to \mathcal{F}''$ over I to **Ab** to be ***exact*** if

$$\mathcal{F}'(U) \to \mathcal{F}(U) \to \mathcal{F}''(U)$$

is an exact sequence for every $U \in I$. If $\mathcal{F}$ is a presheaf on I, define $\mathcal{F}_x$, the ***stalk*** at $x \in X$, by

$$\mathcal{F}_x = \varinjlim_{U \ni x} \mathcal{F}(U).$$

If $\mathcal{F}' \to \mathcal{F} \to \mathcal{F}''$ is an exact sequence of presheaves, prove, for every $x \in X$, that there is an exact sequence of stalks

$$\mathcal{F}'_x \to \mathcal{F}_x \to \mathcal{F}''_x.$$

7.77 (i) Let $F \colon \mathbf{Groups} \to \mathbf{Ab}$ be the functor with $F(G) = G/G'$, where G' is the commutator subgroup of a group G, and let $U \colon \mathbf{Ab} \to \mathbf{Groups}$ be the functor taking every abelian group A into itself (that is, UA regards A as a not necessarily abelian group). Prove that (F, U) is an adjoint pair of functors.

(ii) Prove that the unit of the adjoint pair (F, U) is the natural map $G \to G/G'$.

7.78 Prove that if $T \colon {}_R\mathbf{Mod} \to \mathbf{Ab}$ is an additive left exact functor preserving products, then T preserves inverse limits.

7.79 Generalize Proposition 5.4 to allow infinitely many summands. Let $\{S_i : i \in I\}$ be a family of submodules of an R-module M, where R is a commutative ring. If $M = \langle \bigcup_{i \in I} S_i \rangle$, then the following conditions are equivalent.

(i) $M = \sum_{i \in I} S_i$.

(ii) Every $a \in M$ has a unique expression of the form $a = s_{i_1} + \cdots + s_{i_n}$, where $s_{i_j} \in S_{i_j}$.

(iii) For each $i \in I$,

$$S_i \cap \left\langle \bigcup_{j \neq i} S_j \right\rangle = \{0\}.$$

8

Algebras

This chapter introduces noncommutative rings, along with modules over them. We begin by showing that modules are just another way of viewing representations of rings; that is, ring elements can be viewed as operators on an abelian group. Afterward, we prove the Wedderburn–Artin theorems, which classify semisimple rings, and Maschke's theorem, which says that group algebras are usually semisimple. After a formal interlude investigating tensor products, a construction intimately related to Hom functors (thanks to the *adjoint isomorphism*), we introduce representations and characters of finite groups. This discussion is then applied to prove group-theoretic theorems of Burnside and of Frobenius.

8.1 NONCOMMUTATIVE RINGS

All the rings we have considered so far are commutative, but there are interesting examples of noncommutative rings as well.

Definition. A ***ring*** R is an additive abelian group equipped with a multiplication $R \times R \to R$, denoted by $(a, b) \mapsto ab$, such that, for all $a, b, c \in R$,

(i) $a(bc) = (ab)c$;

(ii) $a(b + c) = ab + ac \quad$ and $\quad (b + c)a = ba + ca$;

(iii) there is $1 \in R$ such that, for all $a \in R$,

$$1a = a = a1.$$

Here are some examples of rings that are not commutative.

Example 8.1.
(i) If k is any commutative ring, then $\mathrm{Mat}_n(k)$, all $n \times n$ matrices with entries in k, is a ring under matrix multiplication and matrix addition; it is commutative if and only if $n = 1$.

If k is not commutative, $\mathrm{Mat}_n(k)$ is a ring, for the usual definition of matrix multiplication still makes sense: If $A = [a_{ij}]$ and $B = [b_{ij}]$, then the ij entry of AB is $\sum_p a_{ip}b_{pj}$; just make sure that entries in A always appear on the left and that entries of B always appear on the right.

(ii) If k is any commutative ring and G is a group (whose operation is written multiplicatively), then we define the ***group algebra*** kG as follows. Its additive abelian group is the free k-module having a basis labeled by the elements of G; thus, each element has a unique expression of the form $\sum_{g\in G} a_g g$, where $a_g \in k$ for all $g \in G$ and *almost all* $a_g = 0$; that is, only finitely many a_g can be nonzero. If g and h are basis elements (i.e., if $g, h \in G$), define their product in kG to be their product gh in G, while $ag = ga$ whenever $a \in k$ and $g \in G$. The product of any two elements of kG is defined by extending by linearity:

$$\Big(\sum_{g\in G} a_g g\Big)\Big(\sum_{h\in G} b_h h\Big) = \sum_{z\in G}\Big(\sum_{gh=z} a_g b_h\Big)z.$$

A group algebra kG is commutative if and only if the group G is abelian.

In Exercise 8.17 on page 533, we give another description of kG, when G is a finite group, as all functions $G \to k$ under pointwise addition and *convolution*.

(iii) An ***endomorphism*** of an abelian group A is a homomorphism $f : A \to A$. The ***endomorphism ring*** of A, denoted by $\mathrm{End}(A)$, is the set of all endomorphisms under pointwise addition

$$f + g : a \mapsto f(a) + g(a),$$

and composition as multiplication. It is easy to check that $\mathrm{End}(A)$ is always a ring, and simple examples show that it may not be commutative. For example, if p is a prime, then $\mathrm{End}(\mathbb{I}_p \oplus \mathbb{I}_p) \cong \mathrm{Mat}_2(\mathbb{F}_p)$.

(iv) Let k be a ring, and let $\sigma : k \to k$ be a ring endomorphism. Define a new multiplication on $k[x] = \{\sum_i a_i x^i : a_i \in k\}$ by

$$xa = \sigma(a)x.$$

Thus, multiplication of two polynomials is now given by

$$\big(\sum_i a_i x^i\big)\big(\sum_j b_j x^j\big) = \sum_r c_r x^r,$$

where $c_r = \sum_{i+j=r} a_i \sigma^i(b_j)$. It is a routine exercise to show that $k[x]$, equipped with this new multiplication, is a not necessarily commutative ring. We denote this ring by $k[x; \sigma]$, and we call it a ring of ***skew polynomials***.

(v) If $R_1, \ldots, R_t$ are rings, then their ***direct product***,

$$R = R_1 \times \cdots \times R_t,$$

is the cartesian product with coordinatewise addition and multiplication:

$$(r_i) + (r_i') = (r_i + r_i') \quad \text{and} \quad (r_i)(r_i') = (r_i r_i');$$

we have abbreviated $(r_1, \ldots, r_t)$ to (r_i).

It is easy to see that $R = R_1 \times \cdots \times R_t$ is a ring. Let us identify $r_i \in R_i$ with the "vector" whose ith coordinate is r_i and whose other coordinates are 0. If $i \neq j$, then $r_i r_j = 0$.

(vi) A ***division ring*** D (or a ***skew field***) is a "noncommutative field"; that is, D is a ring in which $1 \neq 0$ and every nonzero element $a \in D$ has a multiplicative inverse: there exists $a' \in D$ with $aa' = 1 = a'a$. Equivalently, a ring D is a division ring if the set $D^\times$ of its nonzero elements forms a group under multiplication. Of course, fields are division rings; here is a noncommutative example.

Let $\mathbb{H}$ be a four-dimensional vector space over $\mathbb{R}$, and label a basis $1, i, j, k$. Thus, a typical element h in $\mathbb{H}$ is

$$h = a + bi + cj + dk,$$

where $a, b, c, d \in \mathbb{R}$. We define a multiplication of basis elements as follows:

$$i^2 = j^2 = k^2 = -1;$$

$$ij = k = -ji; \qquad jk = i = -kj; \qquad ki = j = -ik,$$

and we insist that every $a \in \mathbb{R}$ commutes with $1, i, j, k$. If we now define a multiplication on arbitrary elements by extending by linearity, then $\mathbb{H}$ is a ring, called the (real) ***quaternions***[1] (associativity of multiplication follows from associativity of multiplication in the group $\mathbf{Q} = \{\pm 1, \pm i, \pm j, \pm k\}$ of quaternions). To see that $\mathbb{H}$ is a division ring, it suffices to find inverses of nonzero elements. Define the ***conjugate*** of $u = a + bi + cj + dk \in \mathbb{H}$ by

$$\overline{u} = a - bi - cj - dk;$$

we see easily that

$$u\overline{u} = a^2 + b^2 + c^2 + d^2.$$

Hence, $u\overline{u} \neq 0$ when $u \neq 0$, and so

$$u^{-1} = \overline{u}/u\overline{u} = \overline{u}/(a^2 + b^2 + c^2 + d^2).$$

It is not difficult to prove that conjugation is an additive isomorphism satisfying

$$\overline{uw} = \overline{w}\,\overline{u}.$$

Just as the Gaussian integers were used to prove Fermat's two-squares theorem (Theorem 3.66)—An odd prime p is a sum of two squares if and only if $p \equiv 1 \bmod 4$—so, too, can the quaternions be used to prove Lagrange's theorem that every positive integer is the sum of four squares (see Samuel, *Algebraic Theory of Numbers*, pages 82–85).

The only property of the field $\mathbb{R}$ we have used in constructing $\mathbb{H}$ is that a sum of nonzero squares be nonzero; any subfield of $\mathbb{R}$ has this property, but $\mathbb{C}$ does not. For example, there is a division ring of rational quaternions.

We shall construct other examples of division rings in Chapter 10 when we discuss *crossed product algebras*. ◄

---

[1]The quaternions were discovered in 1843 by W. R. Hamilton when he was seeking a generalization of the complex numbers to model some physical phenomena. He had hoped to construct a three-dimensional algebra for this purpose, but he succeeded only when he saw that dimension 3 should be replaced by dimension 4. This is why Hamilton called $\mathbb{H}$ the *quaternions*, and this division ring is denoted by $\mathbb{H}$ to honor Hamilton.

Remark. Some mathematicians do not assume, as part of the definition, that rings must contain a unit element 1. They point to natural examples, as the even integers or the integrable functions, where a function $f : [0, \infty) \to \mathbb{R}$ is ***integrable*** if

$$\int_0^\infty |f(x)|\,dx = \lim_{t\to\infty} \int_0^t |f(x)|\,dx < \infty.$$

It is not difficult to see that if f and g are integrable, then so are their pointwise sum $f + g$ and pointwise product fg. The only candidate for a unit is the constant function e with $e(x) = 1$ for all $x \in [0, \infty)$ but, obviously, e is not integrable.

The absence of a unit, however, makes many constructions more complicated. For example, if R is a "ring without unit" and $a \in R$, then defining (a), the principal ideal generated by a, as $(a) = \{ra : r \in R\}$, leads to the possibility that $a \notin (a)$; thus, we must redefine (a) to force a inside. Polynomial rings become strange: If R has no unit, then $x \notin R[x]$. There are other (more important) reasons for wanting a unit, but these examples should suffice to show that not assuming a unit can lead to some awkwardness; therefore, we have decided to insist that rings do have units.

Exercise 8.1 on page 531 shows that every "ring without unit" can be imbedded as an ideal in a ring (with unit). ◂

A *subring* S of a ring R is a ring contained in R so that $1 \in S$ and if $s, s' \in S$, then their sum $s + s'$ and product ss' have the same meaning in S as in R. Here is the formal definition.

Definition. A ***subring*** S of a ring R is a subset of R such that

(i) $1 \in S$;

(ii) if $a, b \in S$, then $a - b \in S$;

(iii) if $a, b \in S$, then $ab \in S$.

Example 8.2.
(i) The ***center*** of a ring R, denoted by $Z(R)$, is the set of all those elements $z \in R$ commuting with everything:

$$Z(R) = \{z \in R : zr = rz \text{ for all } r \in R\}.$$

It is easy to see that $Z(R)$ is a subring of R. If k is a commutative ring, then $k \subseteq Z(kG)$. Exercise 8.10 on page 532 asks you to prove that the center of a matrix ring, $Z(\mathrm{Mat}_n(R))$, is the set of all ***scalar matrices*** aI, where $a \in Z(R)$ and I is the identity matrix; Exercise 8.11 on page 532 says that $Z(\mathbb{H}) = \{a1 : a \in \mathbb{R}\}$.

(ii) If D is a division ring, then its center, $Z(D)$, is a field. Moreover, if $D^\times$ is the multiplicative group of the nonzero elements of D, then $Z(D^\times) = Z(D)^\times$; that is, the center of the multiplicative group $D^\times$ consists of the nonzero elements of $Z(D)$. ◂

Here are two "nonexamples" of subring.

Example 8.3.
(i) Define $S = \{a + ib : a, b \in \mathbb{Z}\} \subseteq \mathbb{C}$. Define addition in S to coincide with addition in $\mathbb{C}$, but define multiplication in S by

$$(a + bi)(c + di) = ac + (ad + bc)i$$

(thus, $i^2 = 0$ in S, whereas $i^2 \neq 0$ in $\mathbb{C}$). It is easy to check that S is a ring, but it is not a subring of $\mathbb{C}$.

(ii) If $R = \mathbb{Z} \times \mathbb{Z}$, then its unit is $(1, 1)$. Let

$$S = \{(n, 0) \in \mathbb{Z} \times \mathbb{Z} : n \in \mathbb{Z}\}.$$

It is easily checked that S is closed under addition and multiplication; indeed, S is a ring, for $(1, 0)$ is the unit in S. However, S is not a subring of R because S does not contain the unit of R. ◄

An immediate complication arising from noncommutativity is that the notion of ideal splinters into three notions. There are now left ideals, right ideals, and two-sided ideals.

Definition. Let R be a ring, and let I be an additive subgroup of R. Then I is a ***left ideal*** if $a \in I$ and $r \in R$ implies $ra \in I$, while I is a ***right ideal*** if $ar \in I$. We say that I is a ***two-sided ideal*** if it is both a left ideal and a right ideal.

Example 8.4.
In $\text{Mat}_2(\mathbb{R})$, the equation

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} u & 0 \\ v & 0 \end{bmatrix} = \begin{bmatrix} * & 0 \\ * & 0 \end{bmatrix}$$

shows that the "first columns" (that is, the matrices that are 0 off the first column), form a left ideal (the "second columns" also form a left ideal.) The equation

$$\begin{bmatrix} u & v \\ 0 & 0 \end{bmatrix} \begin{bmatrix} a & b \\ c & d \end{bmatrix} = \begin{bmatrix} * & * \\ 0 & 0 \end{bmatrix}$$

shows that the "first rows" (that is, the matrices that are 0 off the first row), form a right ideal (the "second rows" also form a right ideal). The reader may show that neither of these one-sided ideals is two-sided; indeed, the only two-sided ideals are $\{0\}$ and $\text{Mat}_2(\mathbb{R})$ itself. This example generalizes, in the obvious way, to give examples of left ideals and of right ideals in $\text{Mat}_n(k)$ for all $n \geq 2$ and every ring k. ◄

Example 8.5.
In a direct product of rings, $R = R_1 \times \cdots \times R_t$, each R_j is identified with

$$R_j = \big\{(0, \ldots, 0, r_j, 0 \ldots, 0) : r_j \in R_j\big\},$$

where r_j occurs in the jth coordinate. It is easy to see that each such R_j is a two-sided ideal in R (for if $j \neq i$, then $r_j r_i = 0$ and $r_i r_j = 0$). Moreover, any left or right ideal in R_j is also a left or right ideal in R. ◄

Homomorphisms $\varphi\colon R \to S$ of rings are defined exactly as in the commutative case; we shall see that their kernels are two-sided ideals. Annihilator ideals, defined in the next section, are another source of two-sided ideals.

Definition. If R and S are rings, then a ***ring homomorphism*** (or *ring map*) is a function $\varphi\colon R \to S$ such that, for all $r, r' \in R$,

(i) $\varphi(r + r') = \varphi(r) + \varphi(r')$;

(ii) $\varphi(rr') = \varphi(r)\varphi(r')$;

(iii) $\varphi(1) = 1$.

If $\varphi\colon R \to S$ is a ring homomorphism, then the ***kernel*** is defined as usual:

$$\ker\varphi = \{r \in R : \varphi(r) = 0\}.$$

The ***image*** is also defined as usual:

$$\operatorname{im}\varphi = \{s \in S : s = \varphi(r) \text{ for some } r \in R\}.$$

The kernel is always a two-sided ideal, for if $\varphi(a) = 0$ and $r \in R$, then

$$\varphi(ra) = \varphi(r)\varphi(a) = 0 = \varphi(a)\varphi(r) = \varphi(ar),$$

so that $a \in \ker\varphi$ implies both ra and ar lie in $\ker\varphi$. On the other hand, $\operatorname{im}\varphi$ is only a subring of S.

We can form the ***quotient ring*** R/I when I is a two-sided ideal, because the multiplication on the quotient abelian group R/I, given by $(r + I)(s + I) = rs + I$, is well-defined: If $r + I = r' + I$ and $s + I = s' + I$, then $rs + I = r's' + I$. That is, if $r - r' \in I$ and $s - s' \in I$, then $rs - r's' \in I$. To see this, note that

$$rs - r's' = rs - rs' + rs' - r's' = r(s - s') + (r - r')s \in I,$$

for both $s - s'$ and $r - r'$ lie in I and, since I is a two-sided ideal, each term on the right side also lies in I. It is easy to see that the ***natural map*** $\pi\colon R \to R/I$, defined (as usual) by $r \mapsto r + I$, is a ring map. It is routine to check that the isomorphism theorems and the correspondence theorem hold for (noncommutative) rings.

We now define R-modules when R is any, not necessarily commutative, ring. In contrast to the commutative case, there are now two different kinds of R-modules: *left R-modules* and *right R-modules*. We have already defined ***left R-modules*** (although we have been calling them *R-modules* until now).

Definition. Let R be a ring. A ***left R-module*** is an (additive) abelian group M equipped with a ***scalar multiplication*** $R \times M \to M$, denoted by

$$(r, m) \mapsto rm,$$

such that the following axioms hold for all $m, m' \in M$ and all $r, r', 1 \in R$:

(i) $r(m + m') = rm + rm'$;

(ii) $(r + r')m = rm + r'm$;

(iii) $(rr')m = r(r'm)$;

(iv) $1m = m$.

Definition. A ***right R-module*** is an (additive) abelian group M equipped with a ***scalar multiplication*** $M \times R \to M$, denoted by

$$(m, r) \mapsto mr,$$

such that the following axioms hold for all $m, m' \in M$ and all $r, r', 1 \in R$:

(i) $(m + m')r = mr + m'r$;

(ii) $m(r + r') = mr + mr'$;

(iii) $m(rr') = (mr)r'$;

(iv) $m1 = m$.

Notation. We denote a left R-module M by ${}_RM$, and we denote a right R-module M by M_R.

Of course, there is nothing to prevent us from denoting the scalar multiplication in a right R-module by $(m, r) \mapsto rm$. If we do so, then we see that only axiom (iii) differs from the axioms for a left R-module; the right version now reads

$$(rr')m = r'(rm).$$

That there is an honest difference between these two definitions is apparent from ideals. A left ideal in a ring R is a left R-module, a right ideal is a right R-module, and we have seen in Example 8.4 that these are different things.

We define ***submodule*** in the obvious way; it is a subgroup that is closed under scalar multiplication. Note that a ring R can be regarded as a left R-module (denoted by ${}_RR$) or as a right R-module (denoted by R_R). The submodules of ${}_RR$ are the left ideals; the submodules of R_R are the right ideals. If N is a submodule of a left R-module M, then the ***quotient module*** M/N is the quotient group made into a left R-module by defining scalar multiplication to be $r(m + N) = rm + N$.

Definition. An additive function $f : M_R \to N_R$ between right R-modules M and N is an ***R-homomorphism*** (or R-map) if $f(mr) = f(m)r$ for all $m \in M$ and $r \in R$. All the right R-modules and R-maps form a category, denoted by $\mathbf{Mod}_R$. The notation ${}_R\mathbf{Mod}$ has already been introduced to denote the category of all left R-modules. In either category, we denote the set of all R-maps between R-modules M and N, where both are R-modules on the same side, by

$$\mathrm{Hom}_R(M, N).$$

Example 8.6.
Let G be a group, let k be a commutative ring, and let A be a left kG-module. Define a new action of G on A, denoted by $g * a$, by

$$g * a = g^{-1}a,$$

where $a \in A$ and $g \in G$. For an arbitrary element of kG, define

$$\Big(\sum_{g \in G} m_g g\Big) * a = \sum_{g \in G} m_g g^{-1} a.$$

It is easy to see that A is a right kG-module under this new action; that is, if $u \in kG$ and $a \in A$, the function $A \times kG \to A$, given by $(a, u) \mapsto u * a$, satisfies the axioms in the definition of right module. Of course, we usually write au instead of $u * a$. Thus, a kG-module can be viewed as either a left or a right kG-module. ◂

Example 8.7.
We now generalize Example 8.1(iii). If M is a left R-module, then an R-map $f : M \to M$ is called an ***R-endomorphism*** of M. The ***endomorphism ring***, denoted by $\mathrm{End}_R(M)$, is the set of all R-endomorphisms of M. As a set, $\mathrm{End}_R(M) = \mathrm{Hom}_R(M, M)$, which we have already seen is an additive abelian group. Now define multiplication to be composition: If $f, g : M \to M$, then $fg : m \mapsto f(g(m))$.

If M is regarded as an abelian group, then we write $\mathrm{End}_{\mathbb{Z}}(M)$ for the endomorphism ring $\mathrm{End}(M)$ (with no subscript) defined in Example 8.1(iii), and $\mathrm{End}_R(M)$ is a subring of $\mathrm{End}_{\mathbb{Z}}(M)$. ◂

We are now going to show that ring elements can be regarded as operators (that is, as endomorphisms) on an abelian group.

Definition. A ***representation*** of a ring R is a ring homomorphism

$$\sigma : R \to \mathrm{End}_{\mathbb{Z}}(M),$$

where M is an abelian group.

Representations of rings can be translated into the language of modules.

Proposition 8.8. *Every representation $\sigma : R \to \mathrm{End}_{\mathbb{Z}}(M)$, where M is an abelian group, equips M with the structure of a left R-module. Conversely, every left R-module M determines a representation $\sigma : R \to \mathrm{End}_{\mathbb{Z}}(M)$.*

Proof. Given a homomorphism $\sigma : R \to \mathrm{End}_{\mathbb{Z}}(M)$, denote $\sigma(r) : M \to M$ by σ_r, and define scalar multiplication $R \times M \to M$ by

$$rm = \sigma_r(m),$$

where $m \in M$. A routine calculation shows that M, equipped with this scalar multiplication, is a left R-module.

Conversely, assume that M is a left R-module. If $r \in R$, then $m \mapsto rm$ defines an endomorphism $T_r : M \to M$. It is easily checked that the function $\sigma : R \to \mathrm{End}_{\mathbb{Z}}(M)$, given by $\sigma : r \mapsto T_r$, is a representation. •

Definition. A left R-module is called ***faithful*** if, for all $r \in R$, whenever $rm = 0$ for all $m \in M$, then $r = 0$.

Of course, M being faithful merely says that the representation $\sigma : R \to \mathrm{End}_{\mathbb{Z}}(M)$ (given in Proposition 8.8) is an injection.

An R-module M is ***finitely generated*** if there are finitely many elements $m_1, \ldots, m_n \in M$ with every $x \in M$ an R-linear combination of $m_1, \ldots, m_n$. In particular, an R-module is ***cyclic*** if it generated by one element.

Example 8.9.
Let E/k be a Galois extension with Galois group $G = \mathrm{Gal}(E/k)$. Then E is a kG-module: If $e \in E$, then

$$\Big(\sum_{\sigma \in G} a_\sigma \sigma\Big)(e) = \sum_{\sigma \in G} a_\sigma \sigma(e).$$

We say that E/k has a ***normal basis*** if E is a cyclic kG-module. Every Galois extension E/k has a normal basis (see Jacobson, *Basic Algebra* I, p. 283). ◄

We can now augment Proposition 7.24, an earlier result about algebraic integers.

Proposition 8.10.

(i) *If M is a finitely generated abelian group that is a faithful left R-module for some ring R, then the additive group of R is finitely generated.*

(ii) *If α is a complex number, let $\mathbb{Z}[\alpha]$ be the subring of $\mathbb{C}$ it generates. If there is a faithful $\mathbb{Z}[\alpha]$-module M that is finitely generated as an abelian group, then α is an algebraic integer.*

Proof. (i) By Proposition 8.8, the ring R is isomorphic to a subring of $\mathrm{End}_{\mathbb{Z}}(M)$. Since M is finitely generated, Exercise 8.6 on page 531 shows that $\mathrm{End}_{\mathbb{Z}}(M) = \mathrm{Hom}_{\mathbb{Z}}(M, M)$ is finitely generated. By Proposition 7.24, the additive group of R is finitely generated.

(ii) By Proposition 7.24, it suffices to prove that the ring $\mathbb{Z}[\alpha]$ is finitely generated as an abelian group, and this follows from part (i). •

We could define right-sided versions of all the previous definitions in Chapter 7—submodule, quotient module, R-homomorphisms, isomorphism theorems, correspondence theorem, direct sums, and so on—but there is a more elegant way to do this.

Definition. Let R be a ring with multiplication $\mu\colon R \times R \to R$. Define the ***opposite ring*** to be the ring R^{op} whose additive group is the same as the additive group of R, but whose multiplication $\mu^{\text{op}}\colon R \times R \to R$ is defined by $\mu^{\text{op}}(r, s) = \mu(s, r) = sr$.

Thus, we have merely reversed the order of multiplication. It is straightforward to check that R^{op} is a ring; it is obvious that $(R^{\text{op}})^{\text{op}} = R$; moreover, $R = R^{\text{op}}$ if and only if R is commutative.

Proposition 8.11. *Every right R-module M is a left R^{op}-module, and every left R-module is a right R^{op}-module.*

Proof. We will be ultra-fussy in this proof. To say that M is a right R-module is to say that there is a function $\sigma : M \times R \to M$, denoted by $\sigma(m, r) = mr$. If $\mu\colon R \times R \to R$ is the given multiplication in R, then axiom (iii) in the definition of right R-module says

$$\sigma(m, \mu(r, r')) = \sigma(\sigma(m, r), r').$$

To obtain a left R-module, define $\sigma'\colon R \times M \to M$ by $\sigma'(r, m) = \sigma(m, r)$. To see that M is a left R^{op}-module, it is only a question of checking axiom (iii), which reads, in the fussy notation,

$$\sigma'(\mu^{\text{op}}(r, r'), m) = \sigma'(r, \sigma'(r', m)).$$

But

$$\sigma'(\mu^{\text{op}}(r, r'), m) = \sigma(m, \mu^{\text{op}}(r, r')) = \sigma(m, \mu(r', r)) = m(r'r),$$

while the right side is

$$\sigma'(r, \sigma'(r', m)) = \sigma(\sigma'(r', m), r) = \sigma(\sigma(m, r'), r) = (mr')r.$$

Thus, the two sides are equal because M is a right R-module.

The second half of the proposition now follows because a right R^{op}-module is a left $(R^{\text{op}})^{\text{op}}$-module; that is, a left R-module. •

It follows from Proposition 8.11 that any theorem about left modules is, in particular, a theorem about left R^{op}-modules, and hence it is also a theorem about right R-modules.

Let us now see that opposite rings are more than an expository device; they do occur in nature.

Proposition 8.12. *If a ring R is regarded as a left module over itself, then there is an isomorphism of rings*

$$\operatorname{End}_R(R) \cong R^{\text{op}}.$$

Proof. Define $\varphi\colon \operatorname{End}_R(R) \to R^{\text{op}}$ by $\varphi(f) = f(1)$; it is routine to check that φ is an isomorphism of additive abelian groups. Now $\varphi(f)\varphi(g) = f(1)g(1)$. On the other hand, $\varphi(fg) = (f \circ g)(1) = f(g(1))$. But if we write $r = g(1)$, then $f(g(1)) = f(r) = f(r \cdot 1) = rf(1)$, because f is an R-map, and so $f(g(1)) = rf(1) = g(1)f(1)$. Therefore,

$$\varphi(fg) = \varphi(g)\varphi(f).$$

We have shown that $\varphi\colon \operatorname{End}_R(R) \to R$ is an additive bijection that reverses multiplication. •

An ***anti-isomorphism*** $\varphi\colon R \to A$, where R and A are rings, is an additive bijection such that

$$\varphi(rs) = \varphi(s)\varphi(r).$$

It is easy to see that R and A are anti-isomorphic if and only if $R \cong A^{\mathrm{op}}$. For example, conjugation in $\mathbb{H}$ is an anti-isomorphism. If k is a commutative ring, then transposition, $A \mapsto A^t$, is an anti-isomorphism $\mathrm{Mat}_n(k) \to \mathrm{Mat}_n(k)$, because $(AB)^t = B^t A^t$; therefore, $\mathrm{Mat}_n(k) \cong [\mathrm{Mat}_n(k)]^{\mathrm{op}}$. However, when k is not commutative, the formula $(AB)^t = B^t A^t$ no longer holds. For example,

$$\left(\begin{bmatrix} a & b \\ c & d \end{bmatrix}\begin{bmatrix} p & q \\ r & s \end{bmatrix}\right)^t = \begin{bmatrix} ap+br & aq+bs \\ cp+dr & cq+ds \end{bmatrix}^t,$$

while

$$\begin{bmatrix} p & q \\ r & s \end{bmatrix}^t \begin{bmatrix} a & b \\ c & d \end{bmatrix}^t = \begin{bmatrix} p & r \\ q & s \end{bmatrix}\begin{bmatrix} a & c \\ b & d \end{bmatrix}$$

has $pa + rb \neq ap + br$ as its 1,1 entry.

Proposition 8.13. *If R is any ring, then*

$$[\mathrm{Mat}_n(R)]^{\mathrm{op}} \cong \mathrm{Mat}_n(R^{\mathrm{op}}).$$

Proof. We claim that transposition $A \mapsto A^t$ is an isomorphism of rings

$$[\mathrm{Mat}_n(R)]^{\mathrm{op}} \to \mathrm{Mat}_n(R^{\mathrm{op}}).$$

First, it follows from $(A^t)^t = A$ that $A \mapsto A^t$ is a bijection. Let us set notation. If $M = [m_{ij}]$ is a matrix, its ij entry m_{ij} may also be denoted by $(M)_{ij}$. Denote the multiplication in R^{op} by $a * b$, where $a * b = ba$, and denote the multiplication in $[\mathrm{Mat}_n(R)]^{\mathrm{op}}$ by $A * B$, where $(A * B)_{ij} = (BA)_{ij} = \sum_k b_{ik}a_{kj} \in R$. We must show that $(A * B)^t = A^t B^t$ in $\mathrm{Mat}_n(R^{\mathrm{op}})$. In $[\mathrm{Mat}_n(R)]^{\mathrm{op}}$, we have

$$\begin{aligned}(A * B)^t_{ij} &= (BA)^t_{ij} \\ &= (BA)_{ji} \\ &= \sum_k b_{jk}a_{ki}.\end{aligned}$$

In $\mathrm{Mat}_n(R^{\mathrm{op}})$, we have

$$\begin{aligned}(A^t B^t)_{ij} &= \sum_k (A^t)_{ik} * (B^t)_{kj} \\ &= \sum_k (A)_{ki} * (B)_{jk} \\ &= \sum_k a_{ki} * b_{jk} \\ &= \sum_k b_{jk}a_{ki}.\end{aligned}$$

Therefore, $(A * B)^t = A^t B^t$ in $\mathrm{Mat}_n(R^{\mathrm{op}})$, as desired. •

Direct sums and direct products of R-modules, where R is any (not necessarily commutative) ring, exist. An R-module is, after all, an additive abelian group equipped with a scalar multiplication. If $\{M_i : i \in I\}$ is a family of left R-modules, construct the ***direct product*** $\prod_{i \in I} M_i$ as the direct product of the underlying abelian groups, and then define scalar multiplication by $r(m_i) = (rm_i)$ if all the M_i are left R-modules, or by $(m_i)r = (m_i r)$ if all the M_i are right R-modules. As with modules over commutative rings, define the ***direct sum*** $\sum_{i \in I} M_i$ as the submodule of $\prod_i M_i$ consisting of all I-tuples almost all of whose coordinates are 0. There is no difficulty in adapting the definition and first properties of external and internal direct sums, such as Proposition 7.15 and Corollary 7.16.

Since direct sums exist, we can also construct ***free*** left R-modules (as direct sums of copies of ${}_R R$) and free right R-modules (as direct sums of R_R).

Exact sequences of left or of right modules also make sense (again, because modules are additive abelian groups with extra structure), and the reader should have no difficulty using them.

EXERCISES

8.1 Let R be an additive abelian group equipped with an associative multiplication that satisfies both distributive laws. Define a multiplication on the abelian group $R^* = \mathbb{Z} \oplus R$ by

$$(m, r)(n, s) = (mn, ms + nr + rs),$$

where ms is the sum of s with itself m times if $m > 0$, and ms is the sum of $-s$ with itself $|m|$ times if $m < 0$.

Prove that R^* is a ring with unit $(1, 0)$, and that R is a two-sided ideal in R^*. (We say that R^* is obtained from R by ***adjoining a unit***.)

8.2 Let R be the set of all matrices of the form $\begin{bmatrix} a & b \\ -\overline{b} & \overline{a} \end{bmatrix}$, where a and b are complex numbers and $\overline{a}$ denotes the complex conjugate of a. Prove that R is a subring of $\mathrm{Mat}_2(\mathbb{C})$ and that $R \cong \mathbb{H}$, where $\mathbb{H}$ is the division ring of quaternions.

8.3 Prove that the following conditions on a ring R are equivalent:

(i) For every sequence of left ideals $L_1 \supseteq L_2 \supseteq L_3 \supseteq \cdots$, there exists N so that $L_i = L_{i+1}$ for all $i \geq N$;

(ii) Every nonempty family $\mathcal{F}$ of left ideals has a minimal element in $\mathcal{F}$.

8.4 (***Change of Rings***) Let $\varphi\colon R \to S$ be a ring homomorphism, and let M be a left S-module. Show that the function $R \times M \to M$, given by $(r, m) \mapsto \varphi(r)m$, defines a scalar multiplication that makes M a left R-module.

8.5 Let I be a two-sided ideal in a ring R. Prove that an abelian group M is a left (R/I)-module if and only if it is a left R-module that is annihilated by I.

8.6 If M is a finitely generated abelian group, prove that the additive group of the ring $\mathrm{End}(M)$ is a finitely generated abelian group.

Hint. There is a finitely generated free abelian group F mapping onto M; apply $\text{Hom}(\ , M)$ to $F \to M \to 0$ to obtain an injection $0 \to \text{Hom}(M, M) \to \text{Hom}(F, M)$. But $\text{Hom}(F, M)$ is a finite direct sum of copies of M.

8.7 (i) If k is a commutative ring and G is a cyclic group of finite order n, prove that $kG \cong k[x]/(x^n - 1)$.

(ii) If k is a domain, define the ring of ***Laurent polynomials*** as the subring of $k(x)$ consisting of all rational functions of the form $f(x)/x^n$ for $n \in \mathbb{Z}$. If G is infinite cyclic, prove that kG is isomorphic to Laurent polynomials.

8.8 Let R be a four-dimensional vector space over $\mathbb{C}$ with basis $1, i, j, k$. Define a multiplication on R so that these basis elements satisfy the same identities satisfied in the quaternions $\mathbb{H}$ [see Example 8.1(vi)]. Prove that R is not a division ring.

8.9 If k is a ring, possibly noncommutative, prove that $\text{Mat}_n(k)$ is a ring.

8.10 Prove that the center of a matrix ring $\text{Mat}_n(R)$ is the set of all scalar matrices aI, where $a \in Z(R)$ and I is the identity matrix.

8.11 Prove that $Z(\mathbb{H}) = \{a1 : a \in \mathbb{R}\}$.

8.12 Let $R = R_1 \times \cdots \times R_m$ be a direct product of rings.

(i) Prove that $R^{\text{op}} = R_1^{\text{op}} \times \cdots \times R_m^{\text{op}}$.

(ii) Prove that $Z(R) = Z(R_1) \times \cdots \times Z(R_m)$.

(iii) If k is a field and

$$R = \text{Mat}_{n_1}(k) \times \cdots \times \text{Mat}_{n_m}(k),$$

prove that $\dim_k(Z(R)) = m$.

8.13 If Δ is a division ring, prove that Δ^{op} is also a division ring.

8.14 An ***idempotent*** in a ring A is an element $e \in A$ with $e^2 = e$. If R is a ring and M is a left R-module, prove that every direct summand $S \subseteq M$ determines an idempotent in $\text{End}_R(M)$.
Hint. See Corollary 7.17.

8.15 Let R be a ring.

(i) (***Peirce Decomposition***). Prove that if e is an idempotent in a ring R, then

$$R = Re \oplus R(1 - e).$$

(ii) Let R be a ring having left ideals I and J such that $R = I \oplus J$. Prove that there are idempotents $e \in I$ and $f \in J$ with $1 = e + f$; moreover, $I = Ie$ and $J = Jf$.

Hint. Decompose $1 = e + f$, and show that $ef = 0 = fe$.

8.16 An element a in a ring R has a ***left inverse*** if there is $u \in R$ with $ua = 1$, and it has a ***right inverse*** if there is $w \in R$ with $aw = 1$.

(i) Prove that if $a \in R$ has both a left inverse u and a right inverse w, then $u = w$.

(ii) Give an example of a ring R in which an element a has two distinct left inverses.

Hint. Define $R = \text{End}_k(V)$, where V is a vector space over a field k with basis $\{b_n : n \geq 1\}$, and define $a \in R$ by $a(b_n) = b_{n+1}$ for all $n \geq 1$.

(iii) (***Kaplansky***) Let R be a ring, and let $a, u, v \in R$ satisfy $ua = 1 = va$. If $u \neq v$, prove that a has infinitely many left inverses.

Hint. Are the elements $u + a^n(1 - au)$ distinct?

8.17 Let k be a field, let G be a finite group, and let $\mathcal{F}(G, k)$ denote the vector space of all functions $G \to k$.

(i) Define $\varphi : kG \to \mathcal{F}(G, k)$ as follows: If $u = \sum_x a_x x \in kG$, then $\varphi_u : x \mapsto a_x$. Prove that

$$\varphi_{u+v} = \varphi_u + \varphi_v$$

and

$$\varphi_{uv}(y) = \sum_{x \in G} \varphi_u(x)\varphi_v(x^{-1}y).$$

(This last operation is called the ***convolution*** of φ_u and φ_v.)

(ii) Prove that $\mathcal{F}(G, k)$ is a ring and that $\Phi\colon kG \to \mathcal{F}(G, k)$, given by $u \mapsto \varphi_u$, is a ring isomorphism.

8.18 (i) For k a field and G a finite group, prove that $(kG)^{\text{op}} \cong kG$.

(ii) Prove that $\mathbb{H}^{\text{op}} \cong \mathbb{H}$, where $\mathbb{H}$ is the division ring of real quaternions.

Exercise 8.30 on page 549 asks for a ring R that is not isomorphic to R^{op}.

8.19 (i) If R is a ring, if $r \in R$, and if $k \subseteq Z(R)$ is a subring, prove that the subring generated by r and k is commutative.

(ii) If Δ is a division ring, if $r \in R$, and if $k \subseteq Z(\Delta)$ is a subring, prove that the subdivision ring generated by r and k is a (commutative) field.

8.20 Write the elements of the group $\mathbf{Q}$ of quaternions as

$$1,\ \overline{1},\ i,\ \overline{i},\ j,\ \overline{j},\ k,\ \overline{k},$$

and define a linear transformation $\varphi : \mathbb{R}\mathbf{Q} \to \mathbb{H}$ by removing the bars:

$$\varphi(\overline{x}) = \varphi(x) = x \quad \text{for } x = 1, i, j, k.$$

Prove that φ is a surjective ring map, and conclude that there is an isomorphism of rings $\mathbb{R}\mathbf{Q}/\ker\varphi \cong \mathbb{H}$. (See Example 9.113 for a less computational proof.)

8.21 If R is a ring in which $x^2 = x$ for every $x \in R$, prove that R is commutative. (A Boolean ring is an example of such a ring.)

8.22 Prove that there is an equivalence of categories ${}_R\mathbf{Mod} \to \mathbf{Mod}_{R^{\text{op}}}$.

Hint. Given a left R-module (M, σ), where M is an additive abelian group and $\sigma\colon R \times M \to M$ is its scalar multiplication, consider the right R^{op}-module (M, σ'), where $\sigma'\colon M \times R^{\text{op}} \to M$ is defined in Proposition 8.11. Define $F\colon {}_R\mathbf{Mod} \to \mathbf{Mod}_{R^{\text{op}}}$ on objects by $(M, \sigma) \mapsto (M, \sigma')$.

8.2 Chain Conditions

This section introduces chain conditions for modules over an arbitrary ring, as well as the Jacobson radical, $J(R)$, a two-sided ideal whose behavior has an impact on a ring R. For example, semisimple rings R are rings that generalize the group ring $\mathbb{C}G$ of a finite group G, and we will characterize them in the next section in terms of $J(R)$ and chain conditions.

We will also prove a theorem of Wedderburn that says that every finite division ring is a field; that is, it is commutative.

We have already proved the Jordan–Hölder theorem for groups (see Theorem 5.52). Here is the version of this theorem for modules. We can prove both of these versions simultaneously if we introduce the notion of *operator groups* (see Robinson, *A Course in the Theory of Groups*, page 65).

Theorem 8.14 (Zassenhaus Lemma). *Given submodules $A \subseteq A^*$ and $B \subseteq B^*$ of a module M (over any ring), there is an isomorphism*

$$\frac{A + (A^* \cap B^*)}{A + (A^* \cap B)} \cong \frac{B + (B^* \cap A^*)}{B + (B^* \cap A)}.$$

Proof. A straightforward adaptation of the proof of Lemma 5.49. •

Definition. A ***series*** (or a ***filtration***) of a module M (over any ring) is a finite sequence of submodules $M = M_0, M_1, M_2, \ldots, M_n = \{0\}$ for which

$$M = M_0 \supseteq M_1 \supseteq M_2 \supseteq \cdots \supseteq M_n = \{0\}.$$

The ***factor modules*** of this series are the modules M_0/M_1, $M_1/M_2, \ldots, M_{n-1}/M_n = M_{n-1}$, and the ***length*** is the number of strict inclusions; equivalently, the length is the number of nonzero factor modules.

A ***refinement*** of a series is a series $M = M'_0, M'_1, \ldots, M'_k = \{0\}$ having the original series as a subsequence. Two series of a module M are ***equivalent*** if there is a bijection between the sets of nonzero factor modules of each so that corresponding factor modules are isomorphic.

Theorem 8.15 (Schreier Refinement Theorem). *Any two series*

$$M = M_0 \supseteq M_1 \supseteq \cdots \supseteq M_n = \{0\} \quad \textit{and} \quad M = N_0 \supseteq N_1 \supseteq \cdots \supseteq N_k = \{0\}$$

of a module M have equivalent refinements.

Proof. A straightforward adaptation of the proof of Theorem 5.51. •

Definition. A left R-module is ***simple*** (or ***irreducible***) if $M \neq \{0\}$ and M has no proper submodules.

As with modules over a commutative ring, the correspondence theorem shows that an R-submodule N of a module M is a maximal submodule if and only if M/N is a simple module. The proof of Corollary 7.14 can be adapted to show that a left R-module S is simple if and only if $S \cong R/I$, where I is a maximal left ideal.

Definition. A ***composition series*** is a series all of whose nonzero factor modules are simple.

Notice that a composition series admits only insignificant refinements; we can merely repeat terms (if M_i/M_{i+1} is simple, then it has no proper nonzero submodules and, hence, there is no intermediate submodule L with $M_i \supsetneq L \supsetneq M_{i+1}$). More precisely, any refinement of a composition series is equivalent to the original composition series.

A module need not have a composition series; for example, the abelian group $\mathbb{Z}$, considered as a $\mathbb{Z}$-module, has no composition series.

Definition. A left R-module M, over any ring R, has the ***ascending chain condition***, abbreviated **ACC**, if every ascending chain of left submodules ***stops***: If

$$S_1 \subseteq S_2 \subseteq S_3 \subseteq \cdots$$

is a chain of submodules, then there is some $t \geq 1$ with

$$S_t = S_{t+1} = S_{t+2} = \cdots .$$

A left R-module M, over any ring R, has the ***descending chain condition***, abbreviated **DCC**, if every descending chain of left submodules ***stops***: If

$$S_1 \supseteq S_2 \supseteq S_3 \supseteq \cdots$$

is a chain of submodules, then there is some $t \geq 1$ with

$$S_t = S_{t+1} = S_{t+2} = \cdots .$$

Most of the theorems proved in Chapter 6 for commutative noetherian rings (for example, Proposition 6.38: The equivalence of the ACC, the maximum condition, and finite generation of ideals) can be generalized, and with the same proofs, to left modules having the ACC.

Proposition 8.16.

(i) *If a left module M has* DCC, *then every nonempty family $\mathcal{F}$ of submodules contains a minimal element*; *that is, there is a submodule $S_0 \in \mathcal{F}$ for which there is no $S \in \mathcal{F}$ with $S \subsetneq S_0$.*

(ii) *If a left module M has* ACC, *then every nonempty family $\mathcal{F}$ of submodules contains a maximal element*; *that is, there is a submodule $S_0 \in \mathcal{F}$ for which there is no $S \in \mathcal{F}$ with $S \supsetneq S_0$.*

Proof. Choose $S \in \mathcal{F}$. If S is a minimal element of $\mathcal{F}$, we are done. Otherwise, there is a submodule $S_1 \in \mathcal{F}$ with $S \supsetneq S_1$. If S_1 is a minimal element, we are done; otherwise, there is a submodule $S_2 \in \mathcal{F}$ with $S \supsetneq S_1 \supsetneq S_2$. The DCC says that this sequence must stop; that is, there is $S_t \in \mathcal{F}$ that is a minimal element of $\mathcal{F}$ (for the only obstruction to finding a smaller submodule is that S_t is minimal). The proof of the second statement is similar. •

Proposition 8.17. *A module M over any ring R has a composition series if and only if it has both chain conditions on submodules.*

Proof. If M has a composition series of length n, then no sequence of submodules can have length $> n$, or we would violate Schreier's theorem (refining a series cannot shorten it). Therefore, M has both chain conditions.

Let $\mathcal{F}_1$ be the family of all the proper submodules of M. By Proposition 8.16, the maximum condition gives a maximal submodule $M_1 \in \mathcal{F}_1$. Let $\mathcal{F}_2$ be the family of all proper submodules of M_1, and let M_2 be maximal such. Iterating, we have a descending sequence

$$M \supsetneq M_1 \supsetneq M_2 \supsetneq \cdots .$$

If M_n occurs in this sequence, the only obstruction to constructing M_{n+1} is if $M_n = 0$. Since M has both chain conditions, this chain must stop, and so $M_t = 0$ for some t. This chain is a composition series of M, for each M_i is a maximal submodule of its predecessor. •

Theorem 8.18 (Jordan–Hölder Theorem). *Any two composition series of a module M are equivalent. In particular, the length of a composition series, if one exists, is an invariant of M, called the* ***length*** *of M.*

Proof. As we remarked earlier, any refinement of a composition series is equivalent to the original composition series. It now follows from Schreier's theorem that any two composition series are equivalent; in particular, they have the same length. •

Let V be a vector space over a field k; if V has dimension n, then V has length n, for if $v_1, \ldots, v_n$ is a basis of V, then a composition series is

$$V = \langle v_1, \ldots, v_n \rangle \supsetneq \langle v_2, \ldots, v_n \rangle \supsetneq \cdots \supsetneq \langle v_n \rangle \supsetneq \{0\}$$

(the factor modules are one-dimensional, and hence are simple k-modules).

Corollary 8.19. *If a module M has length n, then every chain of submodules of M has length $\leq n$.*

Proof. By Schreier's theorem, there is a refinement of the given chain that is a composition series, and so the length of the given chain is at most n. •

The Jordan–Hölder theorem can be regarded as a kind of unique factorization theorem; for example, we saw in Corollary 5.53 that it gives a new proof of the fundamental theorem of arithmetic.

If Δ is a division ring, then a left Δ-module V is called a ***left vector space*** over Δ. The following definition from linear algebra still makes sense here.

Definition. If V is a left vector space over a division ring Δ, then a list $X = x_1, \ldots, x_m$ in V is ***linearly dependent*** if

$$x_i \in \langle x_1, \ldots, \widehat{x_i}, \ldots, x_m \rangle$$

for some i; otherwise, X is called ***linearly independent***.

The reader should check that if $x_1, \ldots, x_m$ is linearly independent, then

$$\langle x_1, \ldots, x_m \rangle = \langle x_1 \rangle \oplus \cdots \oplus \langle x_m \rangle.$$

Proposition 8.20. *Every finitely generated left vector space $V = \langle v_1, \ldots, v_n \rangle$ over a division ring Δ is a direct sum of copies of Δ; that is, every finitely generated left vector space over a division ring has a basis.*

Proof. Consider the series

$$V = \langle v_1, \ldots, v_n \rangle \supseteq \langle v_2, \ldots, v_n \rangle \supseteq \langle v_3, \ldots, v_n \rangle \supseteq \cdots \supseteq \langle v_n \rangle \supseteq \{0\}.$$

Denote $\langle v_{i+1}, \ldots, v_n \rangle$ by U_i, so that $\langle v_i, \ldots, v_n \rangle = \langle v_i \rangle + U_i$. By the second isomorphism theorem,

$$\langle v_i \ldots, v_n \rangle / \langle v_{i+1} \ldots, v_n \rangle = (\langle v_i \rangle + U_i)/U_i \cong \langle v_i \rangle / (\langle v_i \rangle \cap U_i).$$

Therefore, the ith factor module is isomorphic to a quotient of $\langle v_i \rangle \cong \Delta$ if $v_i \neq 0$. Since Δ is a division ring, its only quotients are Δ and $\{0\}$. After throwing away those v_i corresponding to trivial factor modules $\{0\}$, we claim that the remaining v's, denote them by $v_1, \ldots, v_m$, form a basis. For all j, we have $v_j \notin \langle v_{j+1}, \ldots, v_n \rangle$. The reader may now show, by induction on m, that $\langle v_1 \rangle, \ldots, \langle v_m \rangle$ generate a direct sum. •

Another proof of this proposition, using dependency relations, is sketched in Exercise 8.23(ii) on page 548.

The next question is whether any two bases of V have the same number of elements. The proper attitude is that theorems about vector spaces over fields have true analogs for left vector spaces over division rings, but the reader should not merely accept the word of a gentleman and a scholar that this is so.

Corollary 8.21. *If V is a finitely generated left vector space over a division ring Δ, then any two bases of V have the same number of elements.*

Proof. As in the proof of Proposition 8.20, a basis of V gives a series

$$V = \langle v_1, v_2, \ldots, v_n \rangle \supsetneq \langle v_2, \ldots, v_n \rangle \supsetneq \langle v_3, \ldots, v_n \rangle \supsetneq \cdots \supsetneq \langle v_n \rangle \supsetneq \{0\}.$$

This is a composition series, for every factor module is isomorphic to Δ, which is simple because Δ is a division ring. By the Jordan–Hölder theorem, the composition series arising from any other basis of V must have the same length. •

Another proof of this corollary is sketched in Exercise 8.23(iii) on page 548.

It now follows that every finitely generated left vector space V over a division ring Δ has a left dimension, which will be denoted by $\dim(V)$.

If an abelian group V is a left vector space and a right vector space over a division ring Δ, must its left dimension equal its right dimension? There is an example (see Jacobson, *Structure of Rings*, page 158) of a division ring Δ and an abelian group V, which is a vector space over Δ on both sides, with left dimension 2 and right dimension 3.

We have just seen that dimension is well-defined for left vector spaces over division rings. Is the *rank* of a free left R-module F well-defined for every ring R; that is, do any two bases of F have the same number of elements? In Proposition 7.50, we saw that rank is well-defined when R is commutative, and it can be shown that rank is well-defined when R is *left noetherian*; that is, if every left ideal in R is finitely generated (see Rotman, *An Introduction to Homological Algebra*, page 111). However, the next example shows that rank is not always well-defined.

Example 8.22.
Let k be a field, let V be a vector space over k having an infinite basis $\{v_n : n \in \mathbb{N}\}$, and let $R = \text{End}_k(V)$. Let A be the left ideal consisting of all the linear transformations $\varphi\colon V \to V$ for which $\varphi(v_{2n}) = 0$ for all n, and let B be the left ideal consisting of all those linear transformations $\psi\colon V \to V$ for which $\psi(v_{2n+1}) = 0$ for all n. We let the reader check that $A \cap B = \{0\}$ and $A + B = R$, so that $R = A \oplus B$.

Let W be the subspace of V with basis the odd v_{2n+1}. If $f : V \to W$ is a k-isomorphism, then the map $\psi \mapsto f\psi f^{-1}$ is an R-isomorphism

$$R = \text{End}_k(V) \cong \text{End}_k(W) = A.$$

Similarly, if Y is the subspace of V spanned by the even v_{2n}, then $R \cong \text{End}_k(Y) = B$. It follows that the free left R-modules R and $R \oplus R$ are isomorphic. ◀

There is another useful unique factorization theorem. Call a left R-module M, over any ring R, an ***indecomposable*** module if there do not exist nonzero submodules A and B with $M = A \oplus B$. The ***Krull–Schmidt theorem*** says that if M has both chain conditions on submodules, then M is a direct sum of indecomposable modules: $M = A_1 \oplus \cdots \oplus A_n$. Moreover, if $M = B_1 \oplus \cdots \oplus B_m$ is another decomposition into indecomposables, then $m = n$ and there is a permution $\sigma \in S_n$ with $A_i \cong B_{\sigma(i)}$ for all i. A proof can be found in Rotman, *An Introduction to the Theory of Groups*, pages 144–150.

Here is a surprising result of J. M. Wedderburn.

Theorem 8.23 (Wedderburn). *Every finite division ring D is a field*; *that is, multiplication in D is commutative.*

Proof. (***E. Witt***[2]). If Z denotes the center of D, then Z is a finite field, and so it has q elements (where q is a power of some prime). It follows that D is a vector space over Z, and so $|D| = q^n$ for some $n \geq 1$; that is, if we define

$$[D : Z] = \dim_Z(D),$$

then $[D : Z] = n$. The proof will be complete if we can show that $n > 1$ leads to a contradiction.

If $a \in D$, define $C(a) = \{u \in D : ua = au\}$. It is routine to check that $C(a)$ is a subdivision ring of D that contains Z: If $u, v \in D$ commute with a, then so do $u + v$, uv,

[2]We shall give another proof of this in Theorem 9.123.

and u^{-1} (when $u \neq 0$). Consequently, $|C(a)| = q^{d(a)}$ for some integer $d(a)$; that is, $[C(a) : Z] = d(a)$. We do not know whether $C(a)$ is commutative, but Exercise 8.25 on page 548 gives

$$[D : Z] = [D : C(a)][C(a) : Z],$$

where $[D : C(a)]$ denotes the dimension of D as a left vector space over $C(a)$. That is, $n = [D : C(a)]d(a)$, and so $d(a)$ is a divisor of n.

Since D is a division ring, its nonzero elements $D^\times$ form a multiplicative group of order $q^n - 1$. By Example 8.2(ii), the center of the group $D^\times$ is $Z^\times$ and, if $a \in D^\times$, then its centralizer $C_{D^\times}(a) = C(a)^\times$. Hence, $|Z(D^\times)| = q - 1$ and $|C_{D^\times}(a)| = q^{d(a)} - 1$, where $d(a) \mid n$.

The class equation for $D^\times$ is

$$|D^\times| = |Z^\times| + \sum_i [D^\times : C_{D^\times}(a_i)],$$

where one a_i is chosen from each noncentral conjugacy class. But

$$[D^\times : C_{D^\times}(a_i)] = |D^\times|/|C_{D^\times}(a_i)| = (q^n - 1)/(q^{d(a_i)} - 1),$$

so that the class equation becomes

$$q^n - 1 = q - 1 + \sum_i \frac{q^n - 1}{q^{d(a_i)} - 1}. \tag{1}$$

We have already noted that each $d(a_i)$ is a divisor of n, while the condition that a_i is not central says that $d(a_i) < n$.

Recall that the nth cyclotomic polynomial is $\Phi_n(x) = \prod(x - \zeta)$, where ζ ranges over all the primitive nth roots of unity. In Corollary 1.41, we proved that $\Phi_n(q)$ is a common divisor of $q^n - 1$ and $(q^n - 1)/(q^{d(a_i)} - 1)$ for all i, and so Eq. (1) gives

$$\Phi_n(q) \mid (q - 1).$$

If $n > 1$ and ζ is a primitive nth root of unity, then $\zeta \neq 1$, and hence ζ is some other point on the unit circle. Since q is a prime power, it is a point on the x-axis with $q \geq 2$, and so the distance $|q - \zeta| > q - 1$. Therefore,

$$|\Phi_n(q)| = \prod |q - \zeta| > q - 1,$$

and this contradicts $\Phi_n(q) \mid (q - 1)$. We conclude that $n = 1$; that is, $D = Z$, and so D is commutative. •

The next discussion will be used in the next section to prove the Wedderburn–Artin theorems classifying semisimple rings. Let us consider $\text{Hom}_R(A, B)$, where both A and B are left R-modules that are finite direct sums: say, $A = \sum_{i=1}^n A_i$ and $B = \sum_{j=1}^m B_j$. By Theorems 7.32 and 7.33, we have

$$\text{Hom}_R(A, B) \cong \sum_{ij} \text{Hom}_R(A_i, B_j).$$

More precisely, if $\alpha_i : A_i \to A$ is the ith injection and $p_j : B \to B_j$ is the jth projection, then each $f \in \mathrm{Hom}_R(A, B)$ gives maps $f_{ij} = p_j f \alpha_i \in \mathrm{Hom}_R(A_i, B_j)$. Thus, f defines a ***generalized*** $n \times m$ ***matrix*** $[f_{ij}]$ (we call $[f_{ij}]$ a generalized matrix because entries in different positions need not lie in the same algebraic system). The map $f \mapsto [f_{ij}]$ is an isomorphism $\mathrm{Hom}_R(A, B) \to \sum_{ij} \mathrm{Hom}_R(A_i, B_j)$. Similarly, if $g : B \to C$, where $C = \sum_{k=1}^{\ell} C_k$, then g defines a generalized $m \times \ell$ matrix $[g_{jk}]$, where $g_{jk} = q_k g \beta_j : B_j \to C_k$, $\beta_j : B_j \to B$ are the injections, and $q_k : C \to C_k$ are the projections.

The composite $gf : A \to C$ defines a generalized $n \times \ell$ matrix, and we claim that it is given by matrix multiplication $(gf)_{ik} = \sum_j g_{kj} f_{ji}$:

$$\begin{aligned}\sum_j g_{kj} f_{ji} &= \sum_j q_k g \beta_j p_j f \alpha_i \\ &= q_k g \Big(\sum_j \beta_j p_j\Big) f \alpha_i \\ &= q_k g f \alpha_i \\ &= (gf)_{ik},\end{aligned}$$

because $\sum_j \beta_j p_j = 1_B$.

By adding some hypotheses, we can pass from generalized matrices to honest matrices.

Proposition 8.24. *Let $V = \sum_{i=1}^{n} V_i$ be a left R-module. If there is a left R-module L and, for each i, an isomorphism $\varphi_i : V_i \to L$, then there is a ring isomorphism*

$$\mathrm{End}_R(V) \cong \mathrm{Mat}_n(\mathrm{End}_R(L)).$$

Proof. Define

$$\theta : \mathrm{End}_R(V) \to \mathrm{Mat}_n(\mathrm{End}_R(L))$$

by

$$\theta : f \mapsto [\varphi_j p_j f \alpha_i \varphi_i^{-1}],$$

where $\alpha_i : V_i \to V$ and $p_j : V \to V_j$ are injections and projections, respectively. That θ is an additive isomorphism is just the identity

$$\mathrm{Hom}\Big(\sum_i V_i, \sum_i V_i\Big) \cong \sum_{ij} \mathrm{Hom}(V_i, V_j),$$

which holds when the index sets are finite. In the paragraph discussing generalized matrices, the home of the ij entries was $\mathrm{Hom}_R(V_i, V_j)$, whereas the present home of these entries is the isomorphic replica $\mathrm{Hom}_R(L, L) = \mathrm{End}_R(L)$.

We now show that θ preserves multiplication. If $g, f \in \mathrm{End}_R(V)$, then $\theta(gf) =$

$[\varphi_j p_j g f \alpha_i \varphi_i^{-1}]$, while the matrix product is

$$\begin{aligned}\theta(g)\theta(f) &= \Big[\sum_k (\varphi_j p_j g \alpha_k \varphi_k^{-1})(\varphi_k p_k f \alpha_i \varphi_i^{-1})\Big] \\ &= \Big[\sum_k \varphi_j p_j g \alpha_k p_k f \alpha_i \varphi_i^{-1}\Big] \\ &= \Big[\varphi_j p_j g \Big(\sum_k \alpha_k p_k\Big) f \alpha_i \varphi_i^{-1}\Big] \\ &= \Big[\varphi_j p_j g f \alpha_i \varphi_i^{-1}\Big]. \quad \bullet\end{aligned}$$

Corollary 8.25. *If V is an n-dimensional left vector space over a division ring Δ, then there is an isomorphism of rings*

$$\text{End}_\Delta(V) \cong \text{Mat}_n(\Delta)^{\text{op}}.$$

Proof. The isomorphism $\text{End}_k(V) \cong \text{Mat}_n(\Delta^{\text{op}})$ is the special case of Proposition 8.24 for $V = V_1 \oplus \cdots \oplus V_n$, where each V_i is one-dimensional, and hence is isomorphic to Δ. Note that $\text{End}_\Delta(\Delta) \cong \Delta^{\text{op}}$, by Proposition 8.12. Now apply Proposition 8.13, which says that $\text{Mat}_n(\Delta^{\text{op}}) \cong \text{Mat}_n(\Delta)^{\text{op}}$. •

The next result involves a direct sum decomposition at the opposite extreme of that in Proposition 8.24.

Corollary 8.26. *Let an R-module M be a direct sum $M = B_1 \oplus \cdots \oplus B_m$ in which* $\text{Hom}_R(B_i, B_j) = \{0\}$ *for all $i \neq j$. Then there is a ring isomorphism*

$$\text{End}_R(M) \cong \text{End}_R(B_1) \times \cdots \times \text{End}_R(B_m).$$

Proof. If $f, g \in \text{End}_R(M)$, let $[f_{ij}]$ and $[g_{ij}]$ be their generalized matrices. It suffices to show that $[g_{ij}][f_{ij}]$ is the diagonal matrix

$$\text{diag}(g_{11} f_{11}, \ldots, g_{mm} f_{mm}).$$

But if $i \neq j$, then $g_{ik} f_{kj} \in \text{Hom}_R(B_i, B_j) = 0$; hence, $(gf)_{ij} = \sum_k g_{ik} f_{kj} = 0$. •

Definition. If k is a commutative ring, then a ring R is a ***k-algebra*** if R is a k-module and scalars in k commute with everything:

$$a(rs) = (ar)s = r(as)$$

for all $a \in k$ and $r, s \in R$.

If R and S are k-algebras, then a ring homomorphism $f : R \to S$ is called a ***k-algebra map*** if

$$f(ar) = af(r)$$

for all $a \in k$ and $r \in R$; that is, f is also a map of k-modules.

The reason that k is assumed to be commutative (in the definition of k-algebra) can be seen in the important special case when k is a subring of R; setting $s = 1$ and taking $r \in k$ gives $ar = ra$.

Example 8.27.

(i) If $A = \mathbb{C}[x]$, then A is a $\mathbb{C}$-algebra, and $\varphi\colon A \to A$, defined by $\varphi\colon \sum_j c_j x^j \mapsto \sum_j c_j (x-1)^j$ is a $\mathbb{C}$-algebra map. On the other hand, the function $\theta\colon A \to A$, defined by $\theta\colon \sum_j c_j x^j \mapsto \sum_j \overline{c}_j (x-1)^j$ (where $\overline{c}$ is the complex conjugate of c), is a ring map but it is not a $\mathbb{C}$-algebra map. For example, $\theta(ix) = -i(x-1)$ while $i\theta(x) = i(x-1)$. Now $\mathbb{C}[x]$ is also an $\mathbb{R}$-algebra, and θ is an $\mathbb{R}$-algebra map.

(ii) Every ring R is a $\mathbb{Z}$-algebra, and every ring homomorphism is a $\mathbb{Z}$-algebra map. This example shows why, in the definition of R-algebra, we do not demand that k be isomorphic to a subring of R.

(iii) If k is a subring contained in the center of R, then R is a k-algebra.

(iv) If k is a commutative ring, then $\mathrm{Mat}_n(k)$ is a k-algebra.

(v) If k is a commutative ring and G is a group, then the group algebra kG is a k-algebra. ◀

We have already defined the ACC for left modules over any ring. The next definition says that a ring R is left noetherian if it has the ACC when viewed as a left module over itself (recall that its submodules are the left ideals). When R is commutative, this definition specializes to our earlier definition of noetherian ring.

Definition. A ring R is ***left noetherian*** if it has the **ACC** (***ascending chain condition***) on left ideals: every ascending chain of left ideals

$$I_1 \subseteq I_2 \subseteq I_3 \subseteq \cdots$$

stops; that is, there is some $t \geq 1$ with

$$I_t = I_{t+1} = I_{t+2} = \cdots .$$

We define ***right noetherian*** rings similarly as those rings having the ACC on right ideals. If k is a field, then every finite-dimensional k-algebra A is both left and right noetherian, for if $\dim(A) = n$, then there are at most n strict inclusions in any ascending chain of left ideals or of right ideals. In particular, if G is a finite group, then kG is finite-dimensional, and so it is left and right noetherian. Exercise 8.28 on page 549 gives an example of a left noetherian ring that is not right noetherian.

Proposition 8.28. *The following conditions on a ring R are equivalent.*

(i) *R is left noetherian.*

(ii) *Every nonempty family of left ideals of R contains a maximal element.*

(iii) *Every left ideal is finitely generated.*

Proof. Adapt the proof of Proposition 6.38. •

Definition. A ring R is ***left artinian*** if it has the **DCC** (***descending chain condition***): Every descending chain of left ideals

$$I_1 \supseteq I_2 \supseteq I_3 \supseteq \cdots$$

stops; that is, there is some $t \geq 1$ with

$$I_t = I_{t+1} = I_{t+2} = \cdots .$$

We define right artinian rings similarly, and there are examples of left artinian rings that are not right artinian (see Exercise 8.29 on page 549). If k is a field, then every finite-dimensional k-algebra A is both left and right artinian, for if $\dim(A) = n$, then there are at most n strict inclusions in any descending chain of left ideals or of right ideals. In particular, if G is a finite group, then kG is finite-dimensional, and so it is left and right artinian. We conclude that kG has both chain conditions (on both sides) when k is a field and G is a finite group.

The ring $\mathbb{Z}$ is (left) noetherian, but it is not (left) artinian, because the chain

$$\mathbb{Z} \supseteq (2) \supseteq (2^2) \supseteq (2^3) \supseteq \cdots$$

does not stop. In the next section, we will prove that left artinian implies left noetherian.

Definition. A left ideal L in a ring R is a ***minimal left ideal*** if $L \neq \{0\}$ and there is no left ideal J with $\{0\} \subsetneq J \subsetneq L$.

A ring need not contain a minimal left ideal. For example, $\mathbb{Z}$ has no minimal ideals: every nonzero ideal I in $\mathbb{Z}$ has the form $I = (n)$ for some nonzero integer n, and $I = (n) \supsetneq (2n)$.

Proposition 8.29.

(i) *Every minimal left ideal L in a ring R is a simple left R-module.*

(ii) *If R is left artinian, then every nonzero left ideal I contains a minimal left ideal.*

Proof. (i) If L contained a submodule S with $\{0\} \subsetneq S \subsetneq L$, then S would be a left ideal of R, contradicting the minimality of L.

(ii) If $\mathcal{F}$ is the family of all nonzero left ideals contained in I, then $\mathcal{F} \neq \varnothing$ because I is nonzero. By Proposition 8.16, $\mathcal{F}$ has a minimal element, and any such is a minimal left ideal. •

We now define a special ideal, introduced by N. Jacobson, that is the analog of the Frattini subgroup in group theory.

Definition. If R is a ring, then its ***Jacobson radical*** $J(R)$ is defined to be the intersection of all the maximal left ideals in R. A ring R is called ***Jacobson semisimple*** if $J(R) = \{0\}$.

Clearly, we can define another Jacobson radical: the intersection of all the maximal *right* ideals. It turns out, however, that both of these coincide (see Proposition 8.36).

The ring $\mathbb{Z}$ is Jacobson semisimple. The maximal ideals in $\mathbb{Z}$ are the nonzero prime ideals (p), and so $J(\mathbb{Z}) = \bigcap_{p \text{ prime}}(p) = \{0\}$. If R is a *local ring* (a commutative ring having a unique maximal ideal P), then $J(R) = P$. An example of a local ring is $R = \{a/b \in \mathbb{Q} : b \text{ is odd}\}$; its unique maximal ideal is

$$(2) = \{2a/b : b \text{ is odd}\}.$$

Example 8.30.
Let k be a field and let $R = \mathrm{Mat}_n(k)$. For any ℓ between 1 and n, let $\mathrm{COL}(\ell)$ denote the ℓth columns; that is,

$$\mathrm{COL}(\ell) = \big\{A = [a_{ij}] \in \mathrm{Mat}_n(k) : a_{ij} = 0 \text{ for all } j \neq \ell\big\}.$$

It is easy to see that $\mathrm{COL}(\ell) = RE_{\ell\ell}$, where $E_{\ell\ell}$ is the matrix having 1 as its $\ell\ell$ entry and 0s everywhere else. We claim that $\mathrm{COL}(\ell)$ is a minimal left ideal in R. If we define

$$\mathrm{COL}^*(\ell) = \sum_{i \neq \ell} \mathrm{COL}(i),$$

then $\mathrm{COL}^*(\ell)$ is a left ideal with

$$R/\mathrm{COL}^*(\ell) \cong \mathrm{COL}(\ell)$$

as left R-modules. Since $\mathrm{COL}(\ell)$ is a minimal left ideal, it is a simple left R-module, and hence $\mathrm{COL}^*(\ell)$ is a maximal left ideal. Therefore,

$$J(R) \subseteq \bigcap_{\ell} \mathrm{COL}^*(\ell) = \{0\},$$

so that $R = \mathrm{Mat}_n(k)$ is Jacobson semisimple. ◄

Proposition 8.31. *Given a ring R, the following conditions are equivalent for $x \in R$:*

(i) $x \in J(R)$;

(ii) *for every $r \in R$, the element $1 - rx$ has a left inverse; that is, there is $u \in R$ with $u(1 - rx) = 1$;*

(iii) *$x(R/I) = \{0\}$ for every maximal left ideal I (equivalently, $xM = \{0\}$ for every simple left R-module M).*

Proof. (i) $\Rightarrow$ (ii) If there is $r \in R$ with $1 - rx$ not having a left inverse, then $R(1 - rx)$ is a proper left ideal, for it does not contain 1. Hence, there is a maximal left ideal I with $1 - rx \in R(1 - rx) \subseteq I$, for the proof of Theorem 6.46 (Every proper ideal is contained in some maximal ideal) does not use commutativity. Now $rx \in J(R) \subseteq I$, because $J(R)$ is a left ideal, and so $1 = (1 - rx) + rx \in I$, a contradiction.

(ii) $\Rightarrow$ (iii) As we mentioned when simple left R-modules were defined earlier in this chapter, a left R-module M is simple if and only if $M \cong R/I$, where I is a maximal left ideal.

Suppose there is a simple module M for which $xM \neq \{0\}$; hence, there is $m \in M$ with $xm \neq 0$ (of course, $m \neq 0$). It follows that the submodule $Rxm \neq \{0\}$, for it contains $1xm$. Since M is simple, it has only one nonzero submodule, namely, M itself, and so $Rxm = M$. Therefore, there is $r \in R$ with $rxm = m$; that is, $(1 - rx)m = 0$. By hypothesis, $1 - rx$ has a left inverse, say, $u(1 - rx) = 1$. Hence, $0 = u(1 - rx)m = m$, a contradiction.

(iii) $\Rightarrow$ (i) If $x(R/I) = \{0\}$, then $x(1 + I) = x + I = I$; that is, $x \in I$. Therefore, if $x(R/I) = \{0\}$ for every maximal left ideal I, then $x \in \bigcap_I I = J(R)$. •

Notice that condition (ii) in Proposition 8.31 can be restated: $x \in J(R)$ if and only if $1 - z$ has a left inverse for every $z \in Rx$.

The following result is frequently used in commutative algebra.

Corollary 8.32 (Nakayama's Lemma). *If M is a finitely generated left R-module, and if $JM = M$, where $J = J(R)$ is the Jacobson radical, then $M = \{0\}$.*

In particular, if R is a local ring, that is, R is a commutative ring with unique maximal ideal P, and if M is a finitely generated R-module with $PM = M$, then $M = \{0\}$.

Proof. Let $m_1, \ldots, m_n$ be a generating set of M that is minimal in the sense that no proper subset generates M. Since $JM = M$, we have $m_1 = \sum_{i=1}^n r_i m_i$, where $r_i \in J$. It follows that

$$(1 - r_1)m_1 = \sum_{i=2}^n r_i m_i.$$

Since $r_1 \in J$, Proposition 8.31 says that $1 - r_1$ has a left inverse, say, u, and so $m_1 = \sum_{i=2}^n u r_i m_i$. This is a contradiction, for now M can be generated by the proper subset $\{m_2, \ldots, m_n\}$.

The second statement follows at once because $J(R) = P$ when R is a local ring with maximal ideal P. •

Remark. The hypothesis in Nakayama's lemma that the module M be finitely generated is necessary. For example, it is easy to check that $R = \{a/b \in \mathbb{Q} : b \text{ is odd}\}$ is a local ring with maximal ideal $P = (2)$, while $\mathbb{Q}$ is an R-module with $P\mathbb{Q} = 2\mathbb{Q} = \mathbb{Q}$. ◂

Remark. There are other characterizations of $J(R)$. One such will be given in Proposition 8.36, in terms of units in R (elements having two-sided inverses). Another characterization is in terms of *left quasi-regular* elements: An element $x \in R$ is ***left quasi-regular*** if there is $y \in R$ with $y \circ x = 0$ (here, $y \circ x = x + y - yx$ is the *circle operation*), and a left ideal is called ***left quasi-regular*** if each of its elements is left quasi-regular. It can be proved that $J(R)$ is the unique maximal left quasi-regular ideal in R (see Lam, *A First Course in Noncommutative Rings*, pages 67–68). ◄

The next property of an ideal is related to the Jacobson radical.

Definition. A left ideal A in a ring R is ***nilpotent*** if there is some integer $m \geq 1$ with $A^m = \{0\}$.

Recall that A^m is the set of all sums of the form $a_1 \cdots a_m$, where $a_j \in A$ for all j; that is, $A^m = \{\sum_i a_{i1} \cdots a_{im} : a_{ij} \in A\}$. It follows that if A is nilpotent, then every $a \in A$ is nilpotent; that is, $a^m = 0$. On the other hand, if $a \in R$ is a nilpotent element, it does not follow that Ra, the left ideal generated by a, is a nilpotent ideal. For example, let $R = \mathrm{Mat}_2(k)$, for some commutative ring k, and let $a = \left[\begin{smallmatrix} 0 & 1 \\ 0 & 0 \end{smallmatrix}\right]$. Now $a^2 = \left[\begin{smallmatrix} 0 & 0 \\ 0 & 0 \end{smallmatrix}\right]$, but Ra contains

$$e = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix},$$

which is idempotent: $e^2 = e$. Therefore, $e^m = e \neq 0$ for all m, and so $(Re)^m \neq \{0\}$.

Corollary 8.33. *If R is a ring, then $I \subseteq J(R)$ for every nilpotent left ideal I in R.*

Proof. Let $I^n = \{0\}$, and let $x \in I$. For every $r \in R$, we have $rx \in I$, and so $(rx)^n = 0$. The equation

$$(1 + rx + (rx)^2 + \cdots + (rx)^{n-1})(1 - rx) = 1$$

shows that $1 - rx$ is left invertible, and so $x \in J(R)$, by Proposition 8.31. •

Proposition 8.34. *If R is a left artinian ring, then $J(R)$ is a nilpotent ideal.*

Proof. Denote $J(R)$ by J in this proof. The descending chain of left ideals,

$$J \supseteq J^2 \supseteq J^3 \supseteq \cdots,$$

stops, because R is left artinian; say, $J^m = J^{m+1} = \cdots$; define $I = J^m$. It follows that $I = I^2$. We will assume that $I \neq \{0\}$ and reach a contradiction.

Let $\mathcal{F}$ be the family of all nonzero left ideals B with $IB \neq \{0\}$; note that $\mathcal{F} \neq \varnothing$ because $I \in \mathcal{F}$. By Proposition 8.16, there is a minimal element $B_0 \in \mathcal{F}$. Choose $b \in B_0$ with $Ib \neq \{0\}$. Now

$$I(Ib) = I^2 b = Ib \neq \{0\},$$

so that $Ib \subseteq B_0 \in \mathcal{F}$, and minimality gives $B_0 = Ib$. Since $b \in B_0$, there is $x \in I \subseteq J = J(R)$ with $b = xb$. Hence, $0 = (1 - x)b$. But $1 - x$ has a left inverse, say, u, by Proposition 8.31, so that $0 = u(1 - x)b = b$, and this is a contradiction. •

The Jacobson radical is obviously a left ideal, but it turns out to be a right ideal as well; that is, $J(R)$ is a two-sided ideal. We begin by giving another source of two-sided ideals.

Definition. If R is a ring and M is a left R-module, define the ***annihilator*** of M to be

$$\text{ann}(M) = \{a \in R : am = 0 \text{ for all } m \in M\}.$$

Even though it is easy to see that $\text{ann}(M)$ is a two-sided ideal in R, we prove that it is a right ideal. Let $a \in \text{ann}(M)$, $r \in R$, and $m \in M$. Since M is a left R-module, we have $rm \in M$; since a annihilates every element of M, we have $a(rm) = 0$. Finally, associativity gives $(ar)m = 0$ for all m, and so $ar \in \text{ann}(M)$.

Corollary 8.35.

(i) $J(R) = \bigcap_{\substack{I = \text{maximal} \\ \text{left ideal}}} \text{ann}(R/I)$, *and so* $J(R)$ *is a two-sided ideal in* R.

(ii) $R/J(R)$ *is a Jacobson semisimple ring.*

Proof. (i) Let $A(R)$ denote $\bigcap_I \text{ann}(R/I)$, where the intersection is over all maximal left ideals I. For any left ideal I, we claim that $\text{ann}(R/I) \subseteq I$. If $a \in \text{ann}(R/I)$, then, for all $r \in R$, we have $a(r + I) = ar + I = I$; that is, $ar \in I$. In particular, if $r = 1$, then $a \in I$. Hence, $A(R) \subseteq J(R)$.

For the reverse inclusion, assume that I is a maximal left ideal, and define $S = R/I$; maximality of I implies that S is a simple R-module. For each nonzero $x \in S$, define $\varphi_x : R \to S$ by $\varphi_x : r \mapsto rx$. It is easy to check that φ_x is an R-map, and it is surjective because S is simple. Thus, $R/\ker\varphi_x \cong S$, and simplicity of S shows that the left ideal $\ker\varphi_x$ is maximal. But it is easy to see that $\text{ann}(R/I) = \bigcap_{x \in S} \ker\varphi_x$. It follows that $J(R) \subseteq A(R)$. Since $J(R)$ is equal to $A(R)$, which is an intersection of two-sided ideals, $J(R)$ is a two-sided ideal.

(ii) First, $R/J(R)$ is a ring, because $J(R)$ is a two-sided ideal. The correspondence theorem for rings shows that if I is any two-sided ideal of R contained in $J(R)$, then $J(R/I) = J(R)/I$; the result follows if $I = J(R)$. •

Let us now show that we could have defined the Jacobson radical using right ideals instead of left ideals.

Definition. A ***unit*** in a ring R is an element $u \in R$ having a two-sided inverse; that is, there is $v \in R$ with

$$uv = 1 = vu.$$

Proposition 8.36.

(i) *If R is a ring, then*

$$J(R) = \{x \in R : 1 + rxs \text{ is a unit in } R \text{ for all } r, s \in R\}.$$

(ii) *If R is a ring and $J'(R)$ is the intersection of all the maximal right ideals of R, then $J'(R) = J(R)$.*

Proof. (i) Let W be the set of all $x \in R$ such that $1 + rxs$ is a unit for all $r, s \in R$. If $x \in W$, then setting $s = -1$ gives $1 - rx$ a unit for all $r \in R$. Hence, $1 - rx$ has a left inverse, and so $x \in J(R)$, by Proposition 8.31. Therefore, $W \subseteq J(R)$. For the reverse inclusion, let $x \in J(R)$. Since $J(R)$ is a two-sided ideal, by Corollary 8.35, we have $xs \in J(R)$ for all $s \in R$. Proposition 8.31 says that $1 - rxs$ is left invertible for all $r \in R$; that is, there is $u \in R$ with $u(1 - rxs) = 1$. Thus, $u = 1 + urxs$. Now $(-ur)xs \in J(R)$, since $J(R)$ is a two-sided ideal, and so u has a left inverse (Proposition 8.31 once again). On the other hand, u also has a right inverse, namely, $1 - rxs$. By Exercise 8.16, u is a unit in R. Therefore, $1 - rxs$ is a unit in R for all $r, s \in R$. Finally, replacing r by $-r$, we have $1 + rxs$ a unit, and so $J(R) \subseteq W$.

(ii) The description of $J(R)$ in part (i) is left-right symmetric. After proving right-sided versions of Proposition 8.31 and Corollary 8.35, one can see that $J'(R)$ is also described as in part (i). We conclude that $J'(R) = J(R)$. •

EXERCISES

8.23 (i) Generalize the proof of Lemma 6.69 to prove that if Δ is a division ring, then $\alpha \preceq S$, defined by $\alpha \in \langle S \rangle$, is a dependency relation.

(ii) Use Theorem 6.71 to prove that every left vector space over a division ring has a basis.

(iii) Use Theorem 6.72 to prove that any two bases of a left vector space over a division ring have the same cardinality.

8.24 If k is a field and A is a finite-dimensional k-algebra, define

$$L = \{\lambda_a \in \mathrm{End}_k(A) : \lambda_a : x \mapsto ax\}$$

and

$$R = \{\rho_a \in \mathrm{End}_k(A) : \rho_a : x \mapsto xa\}.$$

Prove that there are k-algebra isomorphisms

$$L \cong A \quad \text{and} \quad R \cong A^{\mathrm{op}}.$$

Hint. Show that the function $A \to L$ defined by $a \mapsto \lambda_a$ is an injective k-algebra map which is surjective because A is finite-dimensional.

8.25 (i) Let C be a subdivision ring of a division ring D. Prove that D is a left vector space over C, and conclude that $[D : C] = \dim_C(D)$ is defined.

(ii) If $Z \subseteq C \subseteq D$ is a tower of division rings with $[D : C]$ and $[C : Z]$ finite, then $[D : Z]$ is finite and

$$[D : Z] = [D : C][C : Z].$$

Hint. If $u_1, \ldots, u_m$ is a basis of D as a left vector space over C, and if $c_1, \ldots, c_d$ is a basis of C as a left vector space over Z, show that the set of all $c_i u_j$ (in this order) is a basis of D over Z.

8.26 (***Modular Law***). Let A, B, and A' be submodules of a module M. If $A' \subseteq A$, prove that $A \cap (B + A') = (A \cap B) + A'$.

8.27 (i) Let $0 \to A \to B \to C \to 0$ be an exact sequence of left R-modules over some ring R. Prove that if both A and C have DCC, then B has DCC. Conclude, in this case, that $A \oplus B$ has DCC.

(ii) Let $0 \to A \to B \to C \to 0$ be an exact sequence of left R-modules over some ring R. Prove that if both A and C have ACC, then B has ACC. Conclude, in this case, that $A \oplus B$ has ACC.

(iii) Prove that every semisimple ring is left artinian.

8.28 (***L. Small***) Prove that the ring of all matrices of the form $\left[\begin{smallmatrix} a & 0 \\ b & c \end{smallmatrix}\right]$, where $a \in \mathbb{Z}$ and $b, c \in \mathbb{Q}$, is left noetherian but not right noetherian.

8.29 Let R be the ring of all 2×2 matrices $\left[\begin{smallmatrix} a & b \\ 0 & c \end{smallmatrix}\right]$, where $a \in \mathbb{Q}$ and $b, c \in \mathbb{R}$. Prove that R is right artinian but not left artinian.

Hint. There are only finitely many right ideals in R, but for every $V \subseteq \mathbb{R}$ that is a vector space over $\mathbb{Q}$,

$$\begin{bmatrix} 0 & V \\ 0 & 0 \end{bmatrix} = \left\{ \left[\begin{smallmatrix} 0 & v \\ 0 & 0 \end{smallmatrix}\right] : v \in V \right\}$$

is a left ideal.

8.30 Give an example of a ring R that is not isomorphic to R^{op}.

8.31 (i) If R is a commutative ring with $J(R) = \{0\}$, prove that R has no nilpotent elements.

(ii) Give an example of a commutative ring R having no nilpotent elements and for which $J(R) \neq \{0\}$.

8.32 Let k be a field and $R = \mathrm{Mat}_2(k)$. Prove that $a = \left[\begin{smallmatrix} 0 & 1 \\ 0 & 0 \end{smallmatrix}\right]$ is left quasi-regular, but that the principal left ideal Ra is not a left quasi-regular ideal.

8.33 (i) If Δ is a division ring, prove that a finite subgroup of $\Delta^\times$ need not be cyclic. Compare with Theorem 3.30. (S. A. Amitsur has found all the finite subgroups of multiplicative groups of division rings.)

(ii) If Δ is a division ring whose center is a field of characteristic $p > 0$, prove that every finite subgroup G of $\Delta^\times$ is cyclic.

Hint. Consider $\mathbb{F}_p G$, and use Theorem 8.23.

8.34 If R is a ring and M is a left R-module, prove that $\mathrm{Hom}_R(R, M)$ is a left R-module, and prove that it is isomorphic to M.

Hint. If $f : R \to M$ and $r' \in R$, define $r'f : r \mapsto rr'$.

8.35 If k is a field of characteristic 0, then $\mathrm{End}_k(k[t])$ contains the operators

$$x : f(t) \mapsto \tfrac{d}{dt} f(t) \quad \text{and} \quad y : f(t) \mapsto t f(t).$$

(i) If $A_1(k)$ is the subalgebra of $\mathrm{End}_k(k[t])$ generated by x and y, prove that

$$yx = xy + 1.$$

(ii) Prove that $A_1(k)$ is a left noetherian ring having no proper nontrivial two-sided ideals that satisfies the left and right cancellation laws (if $a \neq 0$, then either equation $ab = ac$ or $ba = ca$ implies $b = c$).

Remark. Exercise 8.35 can be generalized by replacing $k[t]$ by $k[t_1, \ldots, t_n]$, the operator x by partial derivatives

$$x_i\colon f(t_1, \ldots, t_n) \mapsto \tfrac{d}{dt_i} f(t_1, \ldots, t_n),$$

and the operator y by

$$y_i\colon f(t_1, \ldots, t_n) \mapsto t_i f(t_1, \ldots, t_n).$$

The subalgebra $A_n(k)$ of $\mathrm{End}_k(k[t_1, \ldots, t_n])$ generated by $x_1, \ldots, x_n, y_1, \ldots, y_n$ is called the nth ***Weyl algebra*** over k. H. Weyl introduced this algebra to model momentum and position operators in quantum mechanics. It can be shown that $A_n(k)$ is a left noetherian simple domain for all $n \geq 1$ (see McConnell–Robson, *Noncommutative Noetherian Rings*, page 19). ◄

8.3 SEMISIMPLE RINGS

A group is an abstract object; we can picture it only as a "cloud," a capital letter G. Of course, there are familiar concrete groups, such as the symmetric group S_n and the general linear group $\mathrm{GL}(V)$ of all nonsingular linear transformations of a vector space V over a field k. Representations of a finite group G are homomorphisms of G into such familiar groups, and they are of fundamental importance for G.

We begin by showing the connection between group representations and group rings.

Definition. A ***k-representation*** of a group G is a homomorphism

$$\sigma\colon G \to \mathrm{GL}(V),$$

where V is a vector space over a field k.

Note that if $\dim(V) = n$, then $\mathrm{GL}(V)$ contains an isomorphic copy of S_n [if $v_1, \ldots, v_n$ is a basis of V and $\alpha \in S_n$, then there is a nonsingular linear transformation $T\colon V \to V$ with $T(v_i) = v_{\alpha(i)}$ for all i]; therefore, permutation representations are special cases of k-representations. Representations of groups can be translated into the language of kG-modules (compare the next proof with that of Proposition 8.8).

Proposition 8.37. *Every k-representation $\sigma\colon G \to \mathrm{GL}(V)$ equips V with the structure of a left kG-module; denote this module by V^σ. Conversely, every left kG-module V determines a k-representation $\sigma\colon G \to \mathrm{GL}(V)$.*

Proof. Given a homomorphism $\sigma\colon G \to \mathrm{GL}(V)$, denote $\sigma(g)\colon V \to V$ by σ_g, and define an action $kG \times V \to V$ by

$$\Big(\sum_{g \in G} a_g g\Big) v = \sum_{g \in G} a_g \sigma_g(v).$$

A routine calculation shows that V, equipped with this scalar multiplication, is a left kG-module.

Conversely, assume that V is a left kG-module. If $g \in G$, then $v \mapsto gv$ defines a linear transformation $T_g\colon V \to V$; moreover, T_g is nonsingular, for its inverse is $T_{g^{-1}}$. It is easily checked that the function $\sigma\colon G \to \mathrm{GL}(V)$, given by $\sigma\colon g \mapsto T_g$, is a k-representation. •

If $\tau\colon G \to \mathrm{GL}(V)$ is another k-representation, when is $V^\tau \cong V^\sigma$, where V^τ and V^σ are the kG-modules determined by τ, σ, respectively, in Proposition 8.37? Recall that if $T\colon V \to V$ is a linear transformation, then we made V into a $k[x]$-module we denoted by V^T, and we saw, in Proposition 7.3, that if $S\colon V \to V$ is another linear transformation, then $V^S \cong V^T$ if and only if there is a nonsingular $\varphi\colon V \to V$ with $S = \varphi T \varphi^{-1}$.

Proposition 8.38. *Let G be a group and let $\sigma, \tau\colon G \to \mathrm{GL}(V)$ be k-representations, where k is a field. If V^σ and V^τ are the corresponding kG-modules defined in* Proposition 8.37*, then $V^\sigma \cong V^\tau$ as kG-modules if and only if there exists a nonsingular $\varphi\colon V \to V$ with*

$$\varphi\tau(g) = \sigma(g)\varphi$$

for every $g \in G$.

Remark. We often say that φ ***intertwines*** σ and τ. ◂

Proof. If $\varphi\colon V^\tau \to V^\sigma$ is a kG-isomorphism, then $\varphi\colon V \to V$ is an isomorphism of vector spaces with

$$\varphi\left(\sum a_g g v\right) = \left(\sum a_g g\right)\varphi(v)$$

for all $v \in V$ and all $g \in G$. But the definition of scalar multiplication in V^τ is $gv = \tau(g)(v)$, while the definition of scalar multiplication in V^σ is $gv = \sigma(g)(v)$. Hence, for all $g \in G$ and $v \in V$, we have $\varphi(\tau(g)(v)) = \sigma(g)(\varphi(v))$. Therefore,

$$\varphi\tau(g) = \sigma(g)\varphi$$

for all $g \in G$.

Conversely, the hypothesis gives $\varphi\tau(g) = \sigma(g)\varphi$ for all $g \in G$, where φ is a nonsingular k-linear transformation, and so $\varphi(\tau(g)v) = \sigma(g)\varphi(v)$ for all $g \in G$ and $v \in V$. It now follows easily that φ is a kG-isomorphism; that is, φ preserves scalar multiplication by $\sum_{g\in G} a_g g$. •

Let us rephrase the last proposition in terms of matrices.

Corollary 8.39. *Let G be a group and let $\sigma, \tau\colon G \to \mathrm{Mat}_n(k)$ be k-representations. Then $(k^n)^\sigma \cong (k^n)^\tau$ as kG-modules if and only if there is a nonsingular $n \times n$ matrix P with*

$$P\tau(x)P^{-1} = \sigma(x)$$

for every $x \in G$.

Example 8.40.
If G is a finite group and V is a vector space over a field k, then the ***trivial homomorphism*** $\sigma\colon G \to \mathrm{GL}(V)$ is defined by $\sigma(x) = 1_V$ for all $x \in G$. The corresponding kG-module

V^σ is called the ***trivial*** kG-module: If $v \in V$, then $xv = v$ for all $x \in G$. The trivial module k (also called the ***principal*** kG***-module***) is denoted by

$$V_0(k). \quad ◀$$

We now introduce an important class of rings; it will be seen that most group algebras kG are semisimple rings.

Definition. A left R-module is ***semisimple*** if it is a direct sum of simple modules. A ring R is ***left semisimple*** if it is a direct sum of minimal left ideals.[3]

Recall that if a ring R is viewed as a left R-module, then its submodules are its left ideals; moreover, a left ideal is minimal if and only if it is a simple left R-module.

The next proposition generalizes Example 8.30.

Proposition 8.41. *If a ring R is left semisimple, then it has both chain conditions on left ideals.*

Proof. Since R is left semisimple, it is a direct sum of minimal left ideals: $R = \sum_i L_i$. Let $1 = \sum_i e_i$, where $e_i \in L_i$. If $r = \sum_i r_i \in \sum_i L_i$, then $r = 1r$ and so $r_i = e_i r_i$. Hence, if $e_i = 0$, then $L_i = 0$. We conclude that there are only finitely many nonzero L_i; that is, $R = L_1 \oplus \cdots \oplus L_n$. Now the series

$$R = L_1 \oplus \cdots \oplus L_n \supseteq L_2 \oplus \cdots \oplus L_n \supseteq \cdots \supseteq L_n \supseteq \{0\}$$

is a composition series, for the factor modules are $L_1, \ldots, L_n$, which are simple. It follows from Proposition 8.17 that R (as a left R-module over itself) has both chain conditions. •

We now characterize semisimple modules over any ring.

Proposition 8.42. *A left module M (over any ring) is semisimple if and only if every submodule of M is a direct summand.*

Proof. Suppose that M is semisimple; hence, $M = \sum_{j \in J} S_j$, where each S_j is simple. For any subset $I \subseteq J$, define

$$S_I = \sum_{j \in I} S_j.$$

If B is a submodule of M, Zorn's lemma provides a subset $K \subseteq J$ maximal with the property that $S_K \cap B = \{0\}$. We claim that $M = B \oplus S_K$. We must show that $M = B + S_K$, for their intersection is $\{0\}$ by hypothesis, and it suffices to prove that $S_j \subseteq B + S_K$ for

[3] We can define a ring to be ***right semisimple*** if it is a direct sum of minimal right ideals. However, we shall see in Corollary 8.57 that a ring is a left semisimple ring if and only if it is right semisimple.

all $j \in J$. If $j \in K$, then $S_j \subseteq S_K \subseteq B + S_K$. If $j \notin K$, then maximality gives $(S_K + S_j) \cap B \neq \{0\}$. Thus,

$$s_K + s_j = b \neq 0,$$

where $s_K \in S_K$, $s_j \in S_j$, and $b \in B$. Note that $s_j \neq 0$, lest $s_K = b \in S_K \cap B = \{0\}$. Hence,

$$s_j = b - s_K \in S_j \cap (B + S_K),$$

and so $S_j \cap (B + S_K) \neq \{0\}$. But S_j is simple, so that $S_j = S_j \cap (B + S_K)$, and so $S_j \subseteq B + S_K$, as desired. Therefore, $M = B \oplus S_K$.

Now assume that every submodule of M is a direct summand.

(i) Every nonzero submodule B contains a simple summand.

Let $b \in B$ be nonzero. By Zorn's lemma, there exists a submodule C of B maximal with $b \notin C$. By Corollary 7.18, C is a direct summand of B: There is some submodule D with $B = C \oplus D$. We claim that D is simple. If D is not simple, we may repeat the argument just given to show that $D = D' \oplus D''$ for nonzero submodules D' and D''. Thus,

$$B = C \oplus D = C \oplus D' \oplus D''.$$

We claim that at least one of $C \oplus D'$ or $C \oplus D''$ does not contain the original element b. Otherwise, $b = c' + d' = c'' + d''$, where $c', c'' \in C$, $d' \in D'$, and $d'' \in D''$. But $c' - c'' = d'' - d' \in C \cap D = \{0\}$ gives $d' = d'' \in D' \cap D'' = \{0\}$. Hence, $d' = d'' = 0$, and so $b = c' \in C$, contradicting the definition of C. Finally, either $C \oplus D'$ or $C \oplus D''$ contradicts the maximality of C.

(ii) M is left semisimple.

By Zorn's lemma, there is a family $\{S_j : j \in I\}$ of simple submodules of M maximal such that the submodule U they generate is their direct sum: $U = \sum_{j \in I} S_j$. By hypothesis, U is a direct summand: $M = U \oplus V$ for some submodule V of M. If $V = \{0\}$, we are done. Otherwise, by part (i), there is some simple submodule S contained in V that is a summand: $V = S \oplus V'$ for some $V' \subseteq V$. The family $\{S_j : j \in I\} \cup \{S\}$ violates the maximality of the first family of simple submodules, for this larger family also generates its direct sum. Therefore, $V = \{0\}$ and M is left semisimple. •

Corollary 8.43.

(i) *Every submodule and every quotient module of a semisimple module M is itself left semisimple.*

(ii) *If R is a (left) semisimple ring, then every left R-module M is a semisimple module.*

(iii) *If I is a two-sided ideal in a semisimple ring R, then the quotient ring R/I is also a semisimple ring.*

Proof. (i) Let B be a submodule of M. Every submodule C of B is, clearly, a submodule of M. Since M is left semisimple, C is a direct summand of M and so, by Corollary 7.18, C is a direct summand of B. Hence, B is left semisimple, by Proposition 8.42.

Let M/H be a quotient of M. Now H is a direct summand of M, so that $M = H \oplus H'$ for some submodule H' of M. But H' is left semisimple, by the first paragraph, and $M/H \cong H'$.

(ii) There is a free left R-module F and a surjective R-map $\varphi\colon F \to M$. Now R is a semisimple module over itself (this is the definition of semisimple ring), and so F is a semisimple module. Thus, M is a quotient of the semisimple module F, and so it is itself semisimple, by part (i).

(iii) First, R/I is a ring, because I is a two-sided ideal. The left R-module R/I is semisimple, by (i), and so it is a direct sum $R/I \cong \sum S_j$, where the S_j are simple left R-modules. But each S_j is also simple as a left (R/I)-module, for any (R/I)-submodule of S_j is also an R-submodule of S_j. Therefore, R/I is semisimple. •

Corollary 8.44.

(i) *A finitely generated left semisimple R-module M (over a ring R) is a direct sum of a finite number of simple left modules. In particular, a left semisimple ring R is a direct sum of a finite number of minimal left ideals.*

(ii) *The direct product $R = R_1 \times \cdots \times R_m$ of left semisimple rings $R_1, \ldots, R_m$ is also a left semisimple ring.*

Proof. (i) Let $x_1, \ldots, x_n$ be a generating set of M. Since M is left semisimple, it is a direct sum of simple left modules, say, $M = \sum_j S_j$. Now each $x_i = \sum_j s_{ij}$, where $s_{ij} \in S_j$, has only a finite number of nonzero components. Hence, $\{x_1, \ldots, x_n\}$ involves only finitely many S_j's, say, $S_{j_1}, \ldots, S_{j_t}$. Therefore,

$$M \subseteq \langle x_1, \ldots, x_n \rangle \subseteq S_{j_1} \oplus \cdots \oplus S_{j_t} \subseteq M.$$

As a left semisimple module over itself, R is cyclic, hence finitely generated. Therefore, R is a direct sum of only finitely many simple left submodules; that is, R is a direct sum of finitely many minimal left ideals.

(ii) Since each R_i is left semisimple, it is a direct sum of minimal left ideals, say, $R_i = J_{i1} \oplus \cdots \oplus J_{i\,t(i)}$. Each J_{ik} is a left ideal in R, not merely in R_i, as we saw in Example 8.5. It follows that J_{ik} is a minimal left ideal in R. Hence, R is a direct sum of minimal left ideals, and so it is a left semisimple ring. •

It follows that a finite direct product of fields is a commutative semisimple ring (we will prove the converse later in this section). For example, if n is a squarefree integer, then the Chinese remainder theorem implies that $\mathbb{I}_n$ is a semisimple ring. Similarly, if k is a field and $f(x) \in k[x]$ is a product of distinct irreducible polynomials, then $k[x]/(f(x))$ is a semisimple ring.

We can now generalize Proposition 8.20: Every, not necessarily finitely generated, left vector space over a division ring Δ has a basis. Every division ring is a left semisimple ring, and Δ itself is the only minimal left ideal. Therefore, every left Δ-module M is a direct sum of copies of Δ; say, $M = \sum_{i \in I} \Delta_i$. If $x_i \in \Delta_i$ is nonzero, then $X = \{x_i : i \in I\}$ is a basis of M. This observation explains the presence of Zorn's lemma in the proof of Proposition 8.42.

The next result shows that left semisimple rings can be characterized in terms of the Jacobson radical.

Theorem 8.45. *A ring R is left semisimple if and only if it is left artinian and $J(R) = \{0\}$.*

Proof. If R is left semisimple, then there is a left ideal I with $R = J(R) \oplus I$, by Proposition 8.42. It follows from Exercise 8.15(ii) on page 532 that there are idempotents $e \in J(R)$ and $f \in I$ with $1 = e + f$. Since $e \in J(R)$, Proposition 8.31 says that $f = 1 - e$ has a left inverse; there is $u \in R$ with $uf = 1$. But f is an idempotent, so that $f = f^2$. Hence, $1 = uf = uf^2 = (uf)f = f$, so that $e = 1 - f = 0$. Since $J(R)e = J(R)$, by Exercise 8.15(ii) on page 532, we have $J(R) = \{0\}$. Finally, Proposition 8.41 shows that R is left artinian.

Conversely, assume that R is left artinian and $J(R) = \{0\}$. We show first that if I is a minimal left ideal of R, then I is a direct summand of R. Now $I \neq \{0\}$, and so $I \not\subseteq J(R)$; therefore, there is a maximal left ideal A not containing I. Since I is minimal, it is simple, so that $I \cap A$ is either I or $\{0\}$. But $I \cap A = I$ implies $I \subseteq A$, a contradiction, and so $I \cap A = \{0\}$. Maximality of A gives $I + A = R$, and so $R = I \oplus A$.

Choose a minimal left ideal I_1, which exists because R is left artinian. As we have just seen, $R = I_1 \oplus B_1$ for some left ideal B_1. Now B_1 contains a minimal left ideal, say, I_2, by Proposition 8.29(ii), and so there is a left ideal B_2 with $B_1 = I_2 \oplus B_2$. This construction can be iterated to produce a strictly decreasing chain of left ideals $B_1 \supsetneq B_2 \supsetneq \cdots \supsetneq B_{r+1}$ as long as $B_r \neq \{0\}$. If $B_r \neq \{0\}$ for all r, then the DCC is violated. Therefore, $B_r = \{0\}$ for some r, so that $R = I_1 \oplus \cdots \oplus I_r$ and R is semisimple. •

Note that the chain condition is needed. For example, $\mathbb{Z}$ is Jacobson semisimple, that is, $J(\mathbb{Z}) = \{0\}$, but $\mathbb{Z}$ is not a semisimple ring.

We can now prove the following remarkable result.

Theorem 8.46 (Hopkins–Levitzki). *If a ring R is left artinian, then it is left noetherian.*

Proof. It suffices to prove that R, regarded as a left module over itself, has a composition series, for then Proposition 8.17 applies at once to show that R is left noetherian as a module over itself; that is, R has the ACC on left ideals.

If $J = J(R)$ denotes the Jacobson radical, then $J^m = \{0\}$ for some $m \geq 1$, by Proposition 8.34, and so there is a chain

$$R = J^0 \supseteq J \supseteq J^2 \supseteq J^3 \supseteq \cdots \supseteq J^m = \{0\}.$$

Since each J^q is an ideal in R, it has the DCC, as does its quotient J^q/J^{q+1}. Now R/J is a semisimple ring, by Theorem 8.45 [it is left artinian, being a quotient of a left artinian

ring, and Jacobson semisimple, by Corollary 8.35(ii)]. The factor module J^q/J^{q+1} is an (R/J)-module; hence, by Corollary 8.43, J^q/J^{q+1} is a semisimple module, and so it can be decomposed into a direct sum of (possibly infinitely many) simple (R/J)-modules. But there can be only finitely many summands, for every (R/J)-submodule of J^q/J^{q+1} is necessarily an R-submodule, and J^q/J^{q+1} has the DCC on R-submodules. Hence, there are simple (R/J)-modules S_i with

$$J^q/J^{q+1} = S_1 \oplus S_2 \oplus \cdots \oplus S_p.$$

Throwing away one simple summand at a time yields a series of J^q/J^{q+1} whose ith factor module is

$$(S_i \oplus S_{i+1} \oplus \cdots \oplus S_p)/(S_{i+1} \oplus \cdots \oplus S_p) \cong S_i.$$

Now the simple (R/J)-module S_i is also a simple R-module, for it is an R-module annihilated by J, so that we have constructed a composition series for J^q/J^{q+1} as a left R-module. Finally, refine the original series for R in this way, for every q, to obtain a composition series for R. •

Of course, the converse of Theorem 8.46 is false.

The next result is fundamental.

Theorem 8.47 (Maschke's Theorem). *If G is a finite group and k is a field whose characteristic does not divide $|G|$, then kG is a left semisimple ring.*

Remark. The hypothesis always holds if k has characteristic 0. ◂

Proof. By Proposition 8.42, it suffices to prove that every left ideal I of kG is a direct summand. Since k is a field, kG is a vector space over k and I is a subspace. By Corollary 6.49, I is a (vector space) direct summand: There is a subspace V (which may not be a left ideal in kG) with $kG = I \oplus V$. There is a k-linear transformation $d \colon kG \to I$ with $d(b) = b$ for all $b \in I$ and with $\ker d = V$ [each $u \in kG$ has a unique expression of the form $u = b + v$, where $b \in I$ and $v \in V$, and $d(u) = b$]. Were d a kG-map, not merely a k-map, then we would be done, by the criterion of Corollary 7.17: I is a summand of kG if and only if it is a retract; that is, there is a kG-map $D \colon kG \to I$ with $D(u) = u$ for all $u \in I$. We now force d to be a kG-map by an "averaging" process.

Define $D : kG \to kG$ by

$$D(u) = \frac{1}{|G|} \sum_{x \in G} x d(x^{-1}u)$$

for all $u \in kG$. Note that $|G| \neq 0$ in k, by the hypothesis on the characteristic of k, and so it is invertible. It is obvious that D is a k-map.

(i) $\operatorname{im} D \subseteq I$.

If $u \in kG$ and $x \in G$, then $d(x^{-1}u) \in I$ (because $\operatorname{im} d \subseteq I$), and $xd(x^{-1}u) \in I$ because I is a left ideal. Therefore, $D(u) \in I$, for each term in the defining sum of $D(u)$ lies in I.

(ii) If $b \in I$, then $D(b) = b$.

Since $b \in I$, so is $x^{-1}b$, and so $d(x^{-1}b) = x^{-1}b$. Hence, $xd(x^{-1}b) = xx^{-1}b = b$. Therefore, $\sum_{x \in G} xd(x^{-1}b) = |G|b$, and so $D(b) = b$.

(iii) D is a kG-map.

It suffices to prove that $D(gu) = gD(u)$ for all $g \in G$ and all $u \in kG$. But

$$\begin{aligned} gD(u) &= \frac{1}{|G|} \sum_{x \in G} gxd(x^{-1}u) \\ &= \frac{1}{|G|} \sum_{x \in G} gxd(x^{-1}g^{-1}gu) \\ &= \frac{1}{|G|} \sum_{y = gx \in G} yd(y^{-1}gu) \\ &= D(gu) \end{aligned}$$

(as x ranges over all of G, so does $y = gx$). •

The converse of Maschke's theorem is true: If G is a finite group and k is a field whose characteristic p divides $|G|$, then kG is not left semisimple; a proof is outlined in Exercise 8.37 on page 573.

Before analyzing left semisimple rings further, let us give several characterizations of them.

Proposition 8.48. *The following conditions on a ring R are equivalent.*

(i) *R is left semisimple.*

(ii) *Every left R-module is a semisimple module.*

(iii) *Every left R-module is injective.*

(iv) *Every short exact sequence of left R-modules splits.*

(v) *Every left R-module is projective.*

Proof. (i) ⇒ (ii). This follows at once from Corollary 8.43(ii), which says that if R is a semisimple ring, then every R-module is a semisimple module.

(ii) ⇒ (iii). If E is a left R-module, then Proposition 7.64 says that E is injective if every exact sequence $0 \to E \to B \to C \to 0$ splits. By hypothesis, B is a semisimple module, and so Proposition 8.42 implies that the sequence splits; thus, E is injective.

(iii) ⇒ (iv). If $0 \to A \to B \to C \to 0$ is an exact sequence, then it must split because, as every module, A is injective (see Proposition 7.64).

(iv) ⇒ (v). Given a module M, there is an exact sequence

$$0 \to F' \to F \to M \to 0,$$

where F is free. This sequence splits, by hypothesis, and so $F \cong M \oplus F'$. Therefore, M is a direct summand of a free module, and hence it is projective (see Theorem 7.56).

(v) $\Rightarrow$ (i). If I is a left ideal of R, then

$$0 \to I \to R \to R/I \to 0$$

is an exact sequence. By hypothesis, R/I is projective, and so this sequence splits (see Proposition 7.54); that is, I is a direct summand of R. By Proposition 8.42, R is a semisimple left R-module. Therefore, R is a left semisimple ring. •

Modules over semisimple rings are so nice that there is a notion of *global dimension* of a ring R that measures how far removed R is from being semisimple; we will discuss global dimension in Chapter 11.

Here are more examples of left semisimple rings; the Wedderburn–Artin theorem will say that there are no others.

Proposition 8.49.

(i) *If Δ is a division ring and V is a left vector space over Δ with $\dim(V) = n$, then $\mathrm{End}_\Delta(V) \cong \mathrm{Mat}_n(\Delta^{\mathrm{op}})$ is a left semisimple ring.*

(ii) *If $\Delta_1, \ldots, \Delta_m$ are division rings, then*

$$\mathrm{Mat}_{n_1}(\Delta_1) \times \cdots \times \mathrm{Mat}_{n_m}(\Delta_m)$$

is a left semisimple ring.

Proof. (i) By Proposition 8.24, we have

$$\mathrm{End}_\Delta(V) \cong \mathrm{Mat}_n(\mathrm{End}_\Delta(\Delta));$$

by Proposition 8.12, $\mathrm{End}_\Delta(\Delta) \cong \Delta^{\mathrm{op}}$. Therefore, $\mathrm{End}_\Delta(V) \cong \mathrm{Mat}_n(\Delta^{\mathrm{op}})$.

Let us now show that $\mathrm{End}_\Delta(V)$ is semisimple. If $v_1, \ldots, v_n$ is a basis of V, define

$$\mathrm{Col}(j) = \{T \in \mathrm{End}_\Delta(V) : T(v_i) = 0 \text{ for all } i \neq j\}.$$

It is easy to see that $\mathrm{Col}(j)$ is a left ideal in $\mathrm{End}_\Delta(V)$: If $S \in \mathrm{End}_\Delta(V)$, then $S(Tv_i) = 0$ for all $i \neq j$. Recall Example 8.30: If we look in $\mathrm{Mat}_n(\Delta^{\mathrm{op}}) \cong \mathrm{End}_\Delta(V)$, then $\mathrm{Col}(j)$ corresponds to $\mathrm{COL}(j)$, all those matrices whose entries off the jth column are 0. It is obvious that

$$\mathrm{Mat}_n(\Delta^{\mathrm{op}}) = \mathrm{COL}(1) \oplus \cdots \oplus \mathrm{COL}(n).$$

Hence, $\mathrm{End}_\Delta(V)$ is also such a direct sum. We asserted, in Example 8.30, that each $\mathrm{COL}(j)$ is a minimal left ideal, and so $\mathrm{End}_\Delta(V)$ is a left semisimple ring. Let us prove minimality of $\mathrm{Col}(j)$.

Suppose that I is a nonzero left ideal in $\mathrm{End}_\Delta(V)$ with $I \subseteq \mathrm{Col}(j)$. Choose a nonzero $F \in I$; now $F(v_j) = u \neq 0$, for otherwise F would kill every basis element and, hence,

would be 0. If $T \in \mathrm{Col}(j)$, write $T(v_j) = w$. Since $u \neq 0$, there is $S \in \mathrm{End}_\Delta(V)$ with $S(u) = w$. Now

$$SF(v_i) = \begin{cases} 0 & \text{if } i \neq j; \\ S(u) = w & \text{if } i = j. \end{cases}$$

Therefore, $T = SF$, because they agree on a basis, and so $T \in I$, because I is a left ideal. Therefore, $\mathrm{Col}(j) = I$, and $\mathrm{Col}(j)$ is a minimal left ideal.

(ii) This follows at once from part (i) and Proposition 8.44(ii), for if Δ is a division ring, then so is Δ^{op}, by Exercise 8.13 on page 532. •

Corollary 8.50. *If V is an n-dimensional left vector space over a division ring Δ, then the minimal left ideals* $\mathrm{Col}(j)$, *for* $1 \leq j \leq n$, *in* $\mathrm{End}_\Delta(V)$ *are all isomorphic.*

Proof. Let $v_1, \ldots, v_n$ be a basis of V. For each j, define $p_j : V \to V$ to be the linear transformation that interchanges v_j and v_1 and that fixes all the other v_i. It is easy to see that $T \mapsto Tp_j$ is an isomorphism $\mathrm{Col}(1) \to \mathrm{Col}(j)$. •

We will see, in Lemma 8.61(ii), that all the minimal left ideals in $\mathrm{End}_\Delta(V)$ are isomorphic.

Definition. A ring R is ***simple*** if it is nonzero and it has no proper nonzero two-sided ideals.

In Proposition 8.59, we will see that every left artinian simple ring is semisimple.

Proposition 8.51. *If Δ is a division ring, then* $R = \mathrm{Mat}_n(\Delta)$ *is a simple ring.*

Proof. A ***matrix unit*** E_{pq} is the $n \times n$ matrix all of whose entries are 0 except the p, q entry, which is 1. The matrix units form a basis for $\mathrm{Mat}_n(\Delta)$ viewed as a left vector space over Δ, for each matrix $A = [a_{ij}]$ has a unique expression

$$A = \sum_{ij} a_{ij} E_{ij}.$$

[Of course, this says that $\dim(\mathrm{Mat}_n(\Delta)) = n^2$.] A routine calculation shows that matrix units multiply according to the following rule:

$$E_{ij} E_{k\ell} = \begin{cases} 0 & \text{if } j \neq k \\ E_{i\ell} & \text{if } j = k. \end{cases}$$

Suppose that N is a nonzero two-sided ideal in $\mathrm{Mat}_n(\Delta)$. If A is a nonzero matrix in N, it has a nonzero entry; say, $a_{ij} \neq 0$. Since N is a two-sided ideal, N contains $E_{pi} A E_{jq}$

for all p, q. But

$$\begin{aligned} E_{pi}AE_{jq} &= E_{pi}\sum_{k\ell} a_{k\ell}E_{k\ell}E_{jq} \\ &= E_{pi}\sum_{k} a_{kj}E_{kq} \\ &= \sum_{k} a_{kj}E_{pi}E_{kq} \\ &= a_{ij}E_{pq}. \end{aligned}$$

Since $a_{ij} \neq 0$ and Δ is a division ring, $a_{ij}^{-1} \in \Delta$, and so $E_{pq} \in N$ for all p, q. But the collection of all E_{pq} span the left vector space $\mathrm{Mat}_n(\Delta)$ over Δ, and so $N = \mathrm{Mat}_n(\Delta)$. •

We are now going to prove the converse of Proposition 8.49(ii): Every left semisimple ring is isomorphic to a direct product of matrix rings over division rings. The first step shows how division rings arise.

Theorem 8.52 (Schur's Lemma). *Let M and M' be simple left R-modules, where R is a ring.*

(i) *Every nonzero R-map $f : M \to M'$ is an isomorphism.*

(ii) $\mathrm{End}_R(M)$ *is a division ring. In particular, if L is a minimal left ideal in a ring R, then* $\mathrm{End}_R(L)$ *is a division ring.*

Proof. (i) Since M is simple, it has only two submodules: M itself and $\{0\}$. Now the submodule $\ker f \neq M$ because $f \neq 0$, and so $\ker f = \{0\}$; that is, f is an injection. Similarly, the submodule $\mathrm{im}\, f \neq \{0\}$, so that $\mathrm{im}\, f = M'$ and f is a surjection.

(ii) If $f : M \to M$ and $f \neq 0$, then f is an isomorphism, by part (i), and hence it has an inverse $f^{-1} \in \mathrm{End}_R(M)$. Thus, the ring $\mathrm{End}_R(M)$ is a division ring. •

Lemma 8.53. *If L and L' are minimal left ideals in a ring R, then each of the following statements implies the one below it:*

(1) $LL' \neq \{0\}$;

(2) $\mathrm{Hom}_R(L, L') \neq \{0\}$, *and there exists $b' \in L'$ with $L' = Lb'$;*

(3) $L \cong L'$ *as left R-modules.*

If also $L^2 \neq \{0\}$, then (3) *implies* (1), *and the three statements are equivalent.*

Proof. Let L and L' be minimal left ideals.

(1) $\Rightarrow$ (2)

If $LL' \neq \{0\}$, then there exists $b \in L$ and $b' \in L'$ with $bb' \neq 0$. Thus, the function $f : L \to L'$, defined by $x \mapsto xb'$, is a nonzero R-map, and so $\mathrm{Hom}_R(L, L') \neq \{0\}$. Moreover, $Lb' = L'$, for it is a nonzero submodule of the minimal left ideal L'.

(2) $\Rightarrow$ (3)

If $\mathrm{Hom}_R(L, L') \neq \{0\}$, then there is a nonzero $f : L \to L'$, and f is an isomorphism, by Schur's lemma; that is, $L \cong L'$.

(3) and $L^2 \neq \{0\} \Rightarrow$ (1)

Assume now that $L^2 \neq \{0\}$, so there are $x, y \in L$ with $xy \neq 0$. If $g : L \to L'$ is an isomorphism, then $0 \neq g(xy) = xg(y) \in LL'$, and so $LL' \neq \{0\}$. •

Note that if $J(R) = \{0\}$, then $L^2 \neq \{0\}$. Otherwise, L is a nilpotent left ideal and Corollary 8.33 gives $L \subseteq J(R) = \{0\}$, a contradiction.

Proposition 8.54. *If $R = \sum_j L_j$ is a left semisimple ring, where the L_j are minimal left ideals, then every simple R-module S is isomorphic to some L_j.*

Proof. Now $S \cong \mathrm{Hom}_R(R, S) \neq \{0\}$, by Exercise 8.34 on page 549. If $\mathrm{Hom}_R(L_j, S) = \{0\}$ for all j, then $\mathrm{Hom}_R(R, S) = \{0\}$ (for $R = L_1 \oplus \cdots \oplus L_m$). Hence, $\mathrm{Hom}_R(L_j, S) \neq \{0\}$ for some j. Since both L_j and S are simple, Theorem 8.52(i) gives $L_j \cong S$. •

Here is a fancier proof.

Proof. By Corollary 7.14, there is a left ideal I with $S \cong R/I$, and so there is a series

$$R \supseteq I \supseteq \{0\}.$$

In Proposition 8.41, we saw that

$$R = L_1 \oplus \cdots \oplus L_n \supseteq L_2 \oplus \cdots \oplus L_n \supseteq \cdots \supseteq L_n \supseteq \{0\}$$

is a composition series with factor modules $L_1, \ldots, L_n$. The Schreier refinement theorem (Theorem 8.15) now says that these two series have equivalent refinements. Since a composition series admits only refinements that repeat a term, the factor module S occurring in the refinement of the first series must be isomorphic to one of the factor modules in the second series; that is, $S \cong L_i$ for some i. •

Example 8.55.

The trivial kG-module $V_0(k)$ (see Example 8.40) is a simple kG-module (for it is one-dimensional and so has no subspaces other than $\{0\}$ and itself). By Proposition 8.54, $V_0(k)$ is isomorphic to some minimal left ideal L of kG. We shall find L by searching for elements $u = \sum_{g \in G} a_g g$ in kG with $hu = u$ for all $h \in G$. For such elements u,

$$hu = \sum_{g \in G} a_g hg = \sum_{g \in G} a_g g = u.$$

Since the elements in G form a basis for the vector space kG, we may equate coefficients, and so $a_g = a_{hg}$ for all $g \in G$; in particular, $a_1 = a_h$. As this holds for every $h \in G$, all the coefficients a_g are equal. Therefore, if we define $\gamma \in kG$ by

$$\gamma = \sum_{g \in G} g,$$

then u is a scalar multiple of γ. It follows that $L = \langle\gamma\rangle$ is a left ideal isomorphic to the trivial module $V_0(k)$; moreover, $\langle\gamma\rangle$ is the unique such left ideal. ◂

An abstract left semisimple ring R is a direct sum of minimal left ideals: $R = \sum_j L_j$, and we now know that $\mathrm{End}_R(L_j)$ is a division ring for every j. The next step is to find the direct summands of R that will ultimately turn out to be matrix rings; they arise from a decomposition of R into minimal left ideals by collecting isomorphic terms.

Definition. Let R be a left semisimple ring, and let

$$R = L_1 \oplus \cdots \oplus L_n,$$

where the L_j are minimal left ideals. Reindex the summands so that no two of the first m ideals $L_1, \ldots, L_m$ are isomorphic, while every L_j in the given decomposition is isomorphic to some L_i for $1 \le i \le m$. The left ideals

$$B_i = \sum_{L_j \cong L_i} L_j$$

are called the ***simple components*** of R relative to the decomposition $R = \sum_j L_j$.

We shall see, in Corollary 8.62, that the simple components do not depend on the particular decomposition of R as a direct sum of minimal left ideals.

We divide the Wedderburn–Artin[4] theorem into two parts: an existence theorem and a uniqueness theorem.

Theorem 8.56 (Wedderburn–Artin I). *A ring R is left semisimple if and only if R is isomorphic to a direct product of matrix rings over division rings.*

Proof. Sufficiency is Proposition 8.49.

For necessity, assume that R is left semisimple. Now R is the direct sum of its simple components:

$$R = B_1 \oplus \cdots \oplus B_m,$$

where each B_i is a direct sum of isomorphic minimal left ideals. Proposition 8.12 says that there is a ring isomorphism

$$R^{\mathrm{op}} \cong \mathrm{End}_R(R),$$

where R is regarded as a left module over itself. Now $\mathrm{Hom}_R(B_i, B_j) = \{0\}$ for all $i \neq j$, by Lemma 8.53, so that Corollary 8.26 applies to give a ring isomorphism

$$R^{\mathrm{op}} \cong \mathrm{End}_R(R) \cong \mathrm{End}_R(B_1) \times \cdots \times \mathrm{End}_R(B_m).$$

By Proposition 8.24, there is an isomorphism of rings

$$\mathrm{End}_R(B_i) \cong \mathrm{Mat}_{n_i}(\mathrm{End}_R(L_i)),$$

[4]Wedderburn proved the theorem for semisimple k-algebras, where k is a field; Artin generalized the theorem as it is stated here. This theorem is why *artinian* rings are so called.

because B_i is a direct sum of isomorphic copies of L_i. By Schur's lemma, $\text{End}_R(L_i)$ is a division ring, say, Δ_i, and so

$$R^{\text{op}} \cong \text{Mat}_{n_1}(\Delta_1) \times \cdots \times \text{Mat}_{n_m}(\Delta_m).$$

Hence,

$$R \cong [\text{Mat}_{n_1}(\Delta_1)]^{\text{op}} \times \cdots \times [\text{Mat}_{n_m}(\Delta_m)]^{\text{op}}.$$

Finally, Proposition 8.13 gives

$$R \cong \text{Mat}_{n_1}(\Delta_1^{\text{op}}) \times \cdots \times \text{Mat}_{n_m}(\Delta_m^{\text{op}}).$$

This completes the proof, for Δ_i^{op} is also a division ring for all i, by Exercise 8.13 on page 532. •

Corollary 8.57. *A ring R is left semisimple if and only if it is right semisimple.*

Proof. It is easy to see that a ring R is right semisimple if and only if its opposite ring R^{op} is left semisimple. But we saw, in the middle of the proof of Theorem 8.56, that

$$R^{\text{op}} \cong \text{Mat}_{n_1}(\Delta_1) \times \cdots \times \text{Mat}_{n_m}(\Delta_m),$$

where $\Delta_i = \text{End}_R(L_i)$. •

As a consequence of this corollary, we say that a ring is ***semisimple*** without the adjectives left or right.

Corollary 8.58. *A commutative ring R is semisimple if and only if it is isomorphic to a direct product of finitely many fields.*

Proof. A field is a semisimple ring, and so a direct product of finitely many fields is also semisimple, by Corollary 8.44(ii). Conversely, if R is semisimple, it is a direct product of matrix rings over division rings. Since R is commutative, all the matrix rings must be of size 1×1 and all the division rings must be fields. •

Even though the name suggests it, it is not yet clear that a simple ring is semisimple. Indeed, this is false without assuming the DCC (see Lam, *A First Course in Noncommutative Rings*, page 43, for an example of a simple ring that is not semisimple).

Proposition 8.59. *A simple left artinian ring R is semisimple.*

Proof. (***Rieffel***) First, we show that if L is any nonzero left ideal in R and $\Delta = \text{End}_R(L)$, then $R \cong \text{End}_\Delta(L)$. Now L is a left Δ-module [with scalar multiplication $\Delta \times L \to L$ given by $(f, a) \mapsto f(a)$ for all $f \in \Delta$ and $a \in L$].

Define $\varphi \colon R \to \text{End}_\Delta(L)$ by φ_r being left multiplication by r:

$$\varphi_r(a) = ra$$

for all $r \in R$ and $a \in L$. Note that φ_r is a Δ-map: If $f \in \Delta = \mathrm{End}_R(L)$, then

$$\varphi_r(f(a)) = rf(a) = f(ra) = f\varphi_r(a).$$

It is easy to check that φ is a ring homomorphism; in particular, φ_1 is the identity function on L. Since φ is not the zero map, $\ker\varphi \neq R$. But R is a simple ring and $\ker\varphi$ is a two-sided ideal, so that $\ker\varphi = \{0\}$ and φ is an injection.

Proving that φ is a surjection is more subtle. If $b \in L$, define $\rho_b \colon L \to L$ to be right multiplication by b:

$$\rho_b \colon a \mapsto ab.$$

Now $\rho_b \colon L \to L$ is an R-map: If $r \in R$ and $a \in L$, then

$$\rho_b(ra) = (ra)b = r(ab) = r\rho_b(a).$$

Hence, $\rho_b \in \mathrm{End}_R(L) = \Delta$. If $h \in \mathrm{End}_\Delta(L)$ and $a, b \in L$, then

$$h(\rho_b(a)) = \rho_b h(a).$$

The left side is $h(\rho_b(a)) = h(ab) = h(\varphi_a(b))$, and the right side is $\rho_b h(a) = h(a)b = \varphi_{h(a)}(b)$. Therefore,

$$h\varphi_a = \varphi_{h(a)} \in \varphi(L),$$

and so $\varphi(L)$ is a left ideal in $\mathrm{End}_\Delta(L)$.

Now $LR = \{\sum_i v_i r_i \,:\, v_i \in L \text{ and } r_i \in R\}$ is a two-sided ideal in R, and $LR \neq \{0\}$ because R has a unit element. Simplicity of R gives $LR = R$. Therefore, $\varphi(R) = \varphi(LR) = \varphi(L)\varphi(R)$ is a left ideal in $\mathrm{End}_\Delta(L)$ (because $\varphi(L)$ is a left ideal). But $\varphi(R)$ contains $\varphi(1) = 1$, and so the left ideal $\varphi(R)$ contains 1. We conclude that $\varphi(R) = \mathrm{End}_\Delta(L)$ and $R \cong \mathrm{End}_\Delta(L)$.

Since R is left artinian, we may assume that L is a minimal left ideal, that $\Delta = \mathrm{End}_R(L)$ is a division ring (by Schur's lemma), and that L is a left vector space over Δ. If L is finite-dimensional, say, $\dim_\Delta(L) = n$, then $R \cong \mathrm{End}_\Delta(L) \cong \mathrm{Mat}_n(\Delta^{\mathrm{op}})$, and we are done. If, on the other hand, L is infinite-dimensional, then there is an infinite independent set $v_1, v_2, \ldots, v_n, \ldots$ that is part of a basis. If

$$I_j = \{T \in \mathrm{End}_\Delta(L) : T(v_1) = 0 = \cdots = T(v_j)\},$$

then it is easy to see that $I_1 \supsetneq I_2 \supsetneq \cdots$ is a strictly decreasing sequence of left ideals, contradicting R being left artinian. •

The following corollary follows at once from Proposition 8.59 and the Wedderburn–Artin theorem.

Corollary 8.60. *If A is a simple left artinian ring, then $A \cong \mathrm{Mat}_n(\Delta)$ for some $n \geq 1$ and some division ring Δ.*

The next lemma, which gives some interesting properties enjoyed by left semisimple rings, will be used to complete the Wedderburn–Artin theorem by stating uniqueness of its constituent parts. In particular, it will say that the integer n and the division ring Δ in Corollary 8.60 are uniquely determined by A.

Lemma 8.61. *Let R be a left semisimple ring, and let*

$$R = L_1 \oplus \cdots \oplus L_n = B_1 \oplus \cdots \oplus B_m,$$

where the L_j are minimal left ideals and the B_i's are the corresponding simple components of R.

(i) *Each B_i is a ring that is also a two-sided ideal in R, and $B_i B_j = \{0\}$ if $j \neq i$.*

(ii) *If L is any minimal left ideal in R, not necessarily occurring in the given decomposition of R, then $L \cong L_i$ for some i and $L \subseteq B_i$.*

(iii) *Every two-sided ideal D in R is a direct sum of B_i's.*

(iv) *Each B_i is a simple ring.*

Proof. (i) Each B_i is a left ideal. To see that it is also a right ideal, consider

$$B_i R = B_i(B_1 \oplus \cdots \oplus B_m) \subseteq B_i B_1 + \cdots + B_i B_m.$$

Recall, for each i, that B_i is a direct sum of left ideals L isomorphic to L_i. If $L \cong L_i$ and $L' \cong L_j$, then the contrapositive not (3) $\Rightarrow$ not (1) in Lemma 8.53 applies to give $LL' = \{0\}$ if $j \neq i$. Hence, if $j \neq i$,

$$B_i B_j = \Big(\sum_{L \cong L_i} L\Big)\Big(\sum_{L' \cong L_j} L'\Big) \subseteq \sum LL' = \{0\}.$$

Thus, $B_i B_1 + \cdots + B_i B_m \subseteq B_i B_i$. Since B_i is a left ideal, $B_i B_i \subseteq RB_i \subseteq B_i$. Therefore, $B_i R \subseteq B_i$, so that B_i is a right ideal and, hence, is a two-sided ideal.

In the last step, proving that B_i is a right ideal, we saw that $B_i B_i \subseteq B_i$; that is, B_i is closed under multiplication. Therefore, to prove that B_i is a ring, it now suffices to prove that it contains a unit element. If 1 is the unit element in R, then $1 = e_1 + \cdots + e_m$, where $e_i \in B_i$ for all i. If $b_i \in B_i$, then

$$b_i = 1b_i = (e_1 + \cdots + e_m)b_i = e_i b_i,$$

for $B_j B_i = \{0\}$ whenever $j \neq i$, by part (i). Similarly, the equation $b_i = b_i 1$ gives $b_i e_i = b_i$, and so e_i is a unit in B_i. Thus, B_i is a ring.[5]

(ii) By Proposition 8.54, a minimal left ideal L is isomorphic to L_i for some i. Now

$$L = RL = (B_1 \oplus \cdots \oplus B_m)L \subseteq B_1 L + \cdots + B_m L.$$

If $j \neq i$, then $B_j L = \{0\}$, by Lemma 8.53, so that

$$L \subseteq B_i L \subseteq B_i,$$

because B_i is a right ideal.

[5] B_i is not a subring of R because its unit e_i is not the unit 1 in R.

(iii) A nonzero two-sided ideal D in R is a left ideal, and so it contains some minimal left ideal L, by Proposition 8.29(ii). Now $L \cong L_i$ for some i, by Proposition 8.54; we claim that $B_i \subseteq D$. By Lemma 8.53, if L' is any minimal left ideal in B_i, then $L' = Lb'$ for some $b' \in L'$. Since $L \subseteq D$ and D is a right ideal, we have $L' = Lb' \subseteq LL' \subseteq DR \subseteq D$. We have shown that D contains every left ideal isomorphic to L_i; as B_i is generated by such ideals, $B_i \subseteq D$. Write $R = B_I \oplus B_J$, where $B_I = \sum_i B_i$ with $B_i \subseteq D$ and $B_J = \sum_j B_j$ with $B_j \not\subseteq D$. By Corollary 7.18 (which holds for modules over noncommutative rings), $D = B_I \oplus (D \cap B_J)$. But $D \cap B_J = \{0\}$; otherwise, it would contain a minimal left ideal $L \cong L_j$ for some $j \in J$ and, as above, this would force $B_j \subseteq D$. Therefore, $D = B_I$.

(iv) A left ideal in B_i is also a left ideal in R: If $a \in R$, then $a = \sum_j a_j$, where $a_j \in B_j$; if $b_i \in B_i$, then

$$ab_i = (a_1 + \cdots + a_m)b_i = a_i b_i \in B_i,$$

because $B_j B_i = \{0\}$ for $j \neq i$. Similarly, a right ideal in B_i is a right ideal in R, and so a two-sided ideal D in B_i is a two-sided ideal in R. By part (iii), the only two-sided ideals in R are direct sums of simple components, and so $D \subseteq B_i$ implies $D = \{0\}$ or $D = B_i$. Therefore, B_i is a simple ring. •

Corollary 8.62. *If R is a semisimple ring, then the simple component containing a minimal left ideal L_i is the left ideal generated by all the minimal left ideals that are isomorphic to L_i. Therefore, the simple components of a semisimple ring do not depend on a decomposition of R as a direct sum of minimal left ideals.*

Proof. This follows from Lemma 8.61(ii). •

Corollary 8.63.

(i) *If A is a simple artinian ring, then $A \cong \text{Mat}_n(\Delta)$ for some division ring Δ. If L is a minimal left ideal in A, then every simple left A-module is isomorphic to L; moreover, $\Delta^{\text{op}} \cong \text{End}_A(L)$.*

(ii) *Two left A-modules M and N are isomorphic if and only if $\dim_\Delta(M) = \dim_\Delta(N)$. In particular, if $A = \text{Mat}_m(\Delta)$, then $M \cong N$ if and only if $\dim_\Delta(M) = \dim_\Delta(N)$.*

Proof. Since A is a semisimple ring, every left module M is isomorphic to a direct sum of minimal left ideals. But, by Lemma 8.61(ii), all minimal left ideals are isomorphic, say, to L, and so $\dim_\Delta(M)$ is the number of summands in a decomposition. If $M \cong N$ as left $\text{Mat}_n(\Delta)$-modules, then $M \cong N$ as left Δ-modules, and so $\dim_\Delta(M) = \dim_\Delta(N)$. Conversely, if $\dim_\Delta(M) = d = \dim_\Delta(N)$, then both M and N are direct sums of d copies of L, and hence $M \cong N$ as left A-modules.

We may now assume that $A = \text{Mat}_n(\Delta)$ and that $L = \text{Col}(1)$, the minimal left ideal consisting of all the $n \times n$ matrices whose last $n-1$ columns are 0 (see Proposition 8.49). Define $\varphi\colon \Delta \to \text{End}_A(L)$ as follows: if $d \in \Delta$ and $\ell \in L$, then $\varphi_d\colon \ell \mapsto \ell d$. Note that φ_d is an A-map: it is additive and, if $a \in A$ and $\ell \in L$, then $\varphi_d(a\ell) = (a\ell)d = a(\ell d) = a\varphi_d(\ell)$. Next, φ is a ring antihomomorphism: $\varphi_1 = 1_L$, it is additive, and $\varphi_{dd'} = \varphi_{d'}\varphi_d$:

if $\ell \in L$, then $\varphi_{d'}\varphi_d(\ell) = \varphi_d(\ell d') = \ell d' d = \varphi_{dd'}(\ell)$; that is, φ is a ring homomorphism $\Delta^{\text{op}} \to \text{End}_A(L)$. To see that φ is injective, note that each $\ell \in L \subseteq \text{Mat}_n(\Delta)$ is a matrix with entries in Δ; hence, $\ell d = 0$ implies $\ell = 0$. Finally, we show that φ is surjective. Let $f \in \text{End}_A(L)$. Now $L = AE_{11}$, where E_{11} is the matrix unit (every simple module is generated by any nonzero element in it). If $u_i \in \Delta$, let $[u_1, \ldots, u_n]$ denote the $n \times n$ matrix in L whose first column is $(u_1, \ldots, u_n)^t$ and whose other entries are all 0. Write $f(E_{11}) = [d_1, \ldots, d_n]$. If $\ell \in L$, then ℓ has the form $[u_1, \ldots, u_n]$, and using only the definition of matrix multiplication, it is easy to see that $[u_1, \ldots, u_n] = [u_1, \ldots, u_n]E_{11}$. Since f is an A-map,

$$\begin{aligned} f([u_1, \ldots, u_n]) &= f([u_1, \ldots, u_n]E_{11}) \\ &= [u_1, \ldots, u_n] f(E_{11}) \\ &= [u_1, \ldots, u_n][d_1, \ldots, d_n] \\ &= [u_1, \ldots, u_n]d_1 = \varphi_{d_1}([u_1, \ldots, u_n]). \end{aligned}$$

Therefore, $f = \varphi_{d_1} \in \text{im}\,\varphi$, as desired. •

The number m of simple components of R is an invariant, for it is the number of nonisomorphic simple left R-modules. However, there is a much stronger uniqueness result.

Theorem 8.64 (Wedderburn–Artin II). *Every semisimple ring R is a direct product,*

$$R \cong \text{Mat}_{n_1}(\Delta_1) \times \cdots \times \text{Mat}_{n_m}(\Delta_m),$$

where $n_i \geq 1$ and Δ_i is a division ring, and the numbers m and n_i, as well as the division rings Δ_i, are uniquely determined by R.

Proof. Let R be a left semisimple ring, and let $R = B_1 \oplus \cdots \oplus B_m$ be a decomposition into simple components arising from some decomposition of R as a direct sum of minimal left ideals. Suppose that $R = B'_1 \times \cdots \times B'_t$, where each B'_ℓ is a two-sided ideal that is also a simple ring. By Lemma 8.61, each two-sided ideal B'_ℓ is a direct sum of B_i's. But B'_ℓ cannot have more than one summand B_i, lest the simple ring B'_ℓ contain a proper nonzero two-sided ideal. Therefore, $t = m$ and, after reindexing, $B'_i = B_i$ for all i.

Dropping subscripts, it remains to prove that if $B = \text{Mat}_n(\Delta) \cong \text{Mat}_{n'}(\Delta') = B'$, then $n = n'$ and $\Delta \cong \Delta'$. In Proposition 8.49, we proved that $\text{Col}(\ell)$, consisting of the matrices with jth columns 0 for all $j \neq \ell$, is a minimal left ideal in B, so that $\text{Col}(\ell)$ is a simple B-module. Therefore,

$$\{0\} \subseteq \text{Col}(1) \subseteq \text{Col}(1) \oplus \text{Col}(2) \subseteq \cdots \subseteq \text{Col}(1) \oplus \cdots \oplus \text{Col}(n) = B$$

is a composition series of B as a module over itself. By the Jordan-Hölder theorem (Theorem 8.18), n and the factor modules $\text{Col}(\ell)$ are invariants of B. Now $\text{Col}(\ell) \cong \text{Col}(1)$ for all ℓ, by Corollary 8.63, and so it suffices to prove that Δ can be recaptured from $\text{Col}(1)$. But this has been done in Corollary 8.63(i): $\Delta \cong \text{End}_B(\text{Col}(1))^{\text{op}}$. •

The description of the group algebra kG simplifies when the field k is algebraically closed.

Corollary 8.65 (Molien). *If G is a finite group and k is an algebraically closed field whose characteristic does not divide $|G|$, then*

$$kG \cong \mathrm{Mat}_{n_1}(k) \times \cdots \times \mathrm{Mat}_{n_m}(k).$$

Proof. By Maschke's theorem, kG is a semisimple ring, and its simple components are isomorphic to matrix rings of the form $\mathrm{Mat}_n(\Delta)$, where Δ arises as $\mathrm{End}_{kG}(L)^{\mathrm{op}}$ for some minimal left ideal L in kG. Therefore, it suffices to show that $\mathrm{End}_{kG}(L)^{\mathrm{op}} = \Delta = k$.

Now $\mathrm{End}_{kG}(L)^{\mathrm{op}} \subseteq \mathrm{End}_k(L)^{\mathrm{op}}$, which is finite-dimensional over k because L is; hence, $\Delta = \mathrm{End}_{kG}(L)^{\mathrm{op}}$ is finite-dimensional over k. Each $f \in \mathrm{End}_{kG}(L)$ is a kG-map, hence is a k-map; that is, $f(au) = af(u)$ for all $a \in k$ and $u \in L$. Therefore, the map $\varphi_a \colon L \to L$, given by $u \mapsto au$, commutes with f; that is, k (identified with all φ_a) is contained in $Z(\Delta)$, the center of Δ. If $\delta \in \Delta$, then δ commutes with every element in k, and so $k(\delta)$, the subdivision ring generated by k and δ, is a (commutative) field. As Δ is finite-dimensional over k, so is $k(\delta)$; that is, $k(\delta)$ is a finite extension of the field k, and so δ is algebraic over k, by Proposition 3.117. But k is algebraically closed, so that $\delta \in k$ and $\Delta = k$. •

Example 8.66.

There are nonisomorphic finite groups G and H having isomorphic complex group algebras. If G is an abelian group of order n, then $\mathbb{C}G$ is a direct product of matrix rings over $\mathbb{C}$, because $\mathbb{C}$ is algebraically closed. But G abelian implies $\mathbb{C}G$ commutative. Hence, $\mathbb{C}G$ is the direct product of n copies of $\mathbb{C}$. It follows that if H is any abelian group of order n, then $\mathbb{C}G \cong \mathbb{C}H$. In particular, $\mathbb{I}_4$ and $\mathbb{I}_2 \oplus \mathbb{I}_2$ are nonisomorphic groups having isomorphic complex group algebras. It follows from this example that certain properties of a group G get lost in the group algebra $\mathbb{C}G$. ◂

Corollary 8.67. *If G is a finite group and k is an algebraically closed field whose characteristic does not divide $|G|$, then $|G| = n_1^2 + n_2^2 + \cdots + n_m^2$, where the ith simple component B_i of kG consists of $n_i \times n_i$ matrices. Moreover, we may assume that $n_1 = 1$.*[6]

Remark. Theorem 8.149 says that all the n_i are divisors of $|G|$. ◂

Proof. As vector spaces over k, both kG and $\mathrm{Mat}_{n_1}(k) \times \cdots \times \mathrm{Mat}_{n_m}(k)$ have the same dimension, for they are isomorphic, by Corollary 8.65. But $\dim(kG) = |G|$, and the dimension of the right side is $\sum_i \dim(\mathrm{Mat}_{n_i}(k)) = \sum_i n_i^2$.

Finally, Example 8.55 shows that there is a unique minimal left ideal isomorphic to the trivial module $V_0(k)$; the corresponding simple component, say, B_1, is one-dimensional, and so $n_1 = 1$. •

The number m of simple components in $\mathbb{C}G$ has a group-theoretic interpretation; we begin by finding the center of the group algebra.

Definition. Let $C_1, \ldots, C_r$ be the conjugacy classes in a finite group G. For each C_j, define the ***class sum*** to be the element $z_j \in \mathbb{C}G$ given by

$$z_j = \sum_{g \in C_j} g.$$

[6]By Example 8.55, the group algebra kG always has a unique minimal left ideal isomorphic to $V_0(k)$, even when k is not algebraically closed.

Here is a ring-theoretic interpretation of the number c of conjugacy classes.

Lemma 8.68. *If r is the number of conjugacy classes in a finite group G, then*

$$r = \dim_{\mathbb{C}}(Z(\mathbb{C}G)),$$

where $Z(\mathbb{C}G))$ is the center of the group algebra. In fact, a basis of $Z(\mathbb{C}G)$ consists of all the class sums.

Proof. If $z_j = \sum_{g \in C_j} g$ is a class sum, then we claim that $z_j \in Z(\mathbb{C}G)$. If $h \in G$, then $hz_jh^{-1} = z_j$, because conjugation by any element of G merely permutes the elements in a conjugacy class. Note that if $j \neq \ell$, then z_j and z_ℓ have no nonzero components in common, and so $z_1, \ldots, z_r$ is a linearly independent list. It remains to prove that the z_j span the center.

Let $u = \sum_{g \in G} a_g g \in Z(\mathbb{C}G)$. If $h \in G$, then $huh^{-1} = u$, and so $a_{hgh^{-1}} = a_g$ for all $g \in G$. Thus, if g and g' lie in the same conjugacy class of G, then their coefficients in u are the same. But this says that u is a linear combination of the class sums z_j. •

Theorem 8.69. *If G is a finite group, then the number m of simple components in $\mathbb{C}G$ is equal to the number r of conjugacy classes in G.*

Proof. We have just seen, in Lemma 8.68, that $r = \dim_{\mathbb{C}}(Z(\mathbb{C}G))$. On the other hand, $Z(\mathrm{Mat}_{n_i}(\mathbb{C}))$, the center of a matrix ring, is the subspace of all scalar matrices, so that $m = \dim_{\mathbb{C}}(Z(\mathbb{C}G))$, by Exercise 8.12(iii) on page 532. •

We began this section by seeing that k-representations of a group G correspond to kG-modules. Let us now return to representations.

Definition. A k-representation of a group G is ***irreducible*** if the corresponding kG-module is simple.

For example, a one-dimensional (necessarily irreducible) k-representation is a group homomorphism $\lambda \colon G \to k^\times$, where $k^\times$ is the multiplicative group of nonzero elements of k. The trivial kG-module $V_0(k)$ corresponds to the representation $\lambda_g = 1$ for all $g \in G$.

The next result is basic to the construction of the character table of a finite group.

Theorem 8.70. *If G is a finite group, then the number of its irreducible complex representations is equal to the number r of its conjugacy classes.*

Proof. By Proposition 8.54, every simple $\mathbb{C}G$-module is isomorphic to a minimal left ideal. Since the number of minimal left ideals is m [the number of simple components of $\mathbb{C}G$], we see that m is the number of irreducible $\mathbb{C}$-representations of G. But Theorem 8.69 equates m with the number r of conjugacy classes in G. •

Example 8.71.
(i) If $G = S_3$, then $\mathbb{C}G$ is six-dimensional. There are three simple components, for S_3 has three conjugacy classes (by Theorem 2.9, the number of conjugacy classes in S_n is equal to the number of different cycle structures), having dimensions 1, 1, and 4, respectively. (We could have seen this without Theorem 8.69, for this is the only way to write 6 as a sum of squares aside from a sum of six 1's.) Therefore,

$$\mathbb{C}S_3 \cong \mathbb{C} \times \mathbb{C} \times \mathrm{Mat}_2(\mathbb{C}).$$

One of the one-dimensional irreducible representations is the trivial one; the other is sgn (signum).

(ii) We now analyze kG for $G = \mathbf{Q}$, the quaternion group of order 8. If $k = \mathbb{C}$, then Corollary 8.65 gives

$$\mathbb{C}\mathbf{Q} \cong \mathrm{Mat}_{n_1}(\mathbb{C}) \times \cdots \times \mathrm{Mat}_{n_r}(\mathbb{C}),$$

while Corollary 8.67 gives

$$|\mathbf{Q}| = 8 = n_1^2 + n_2^2 + \cdots + n_r^2,$$

where $n_1 = 1$. It follows that either all $n_i = 1$ or four $n_i = 1$ and one $n_i = 2$. The first case cannot occur, for it would imply that $\mathbb{C}\mathbf{Q}$ is a commutative ring, whereas the group $\mathbf{Q}$ of quaternions is not abelian. Therefore,

$$\mathbb{C}\mathbf{Q} \cong \mathbb{C} \times \mathbb{C} \times \mathbb{C} \times \mathbb{C} \times \mathrm{Mat}_2(\mathbb{C}).$$

We could also have used Theorem 8.69, for $\mathbf{Q}$ has exactly five conjugacy classes, namely, $\{1\}$, $\{\overline{1}\}$, $\{i, \overline{i}\}$, $\{j, \overline{j}\}$, $\{k, \overline{k}\}$.

The group algebra $\mathbb{R}\mathbf{Q}$ is more complicated because $\mathbb{R}$ is not algebraically closed. Exercise 8.20 on page 533 shows that $\mathbb{H}$ is a quotient of $\mathbb{R}\mathbf{Q}$, hence $\mathbb{H}$ is isomorphic to a direct summand of $\mathbb{R}\mathbf{Q}$ because $\mathbb{R}\mathbf{Q}$ is semisimple. It turns out that

$$\mathbb{R}\mathbf{Q} \cong \mathbb{R} \times \mathbb{R} \times \mathbb{R} \times \mathbb{R} \times \mathbb{H}. \quad ◀$$

Here is an amusing application of the Wedderburn–Artin theorems.

Proposition 8.72. *Let R be a ring whose group of units $U = U(R)$ is finite and of odd order. Then U is abelian and there are positive integers m_i with*

$$|U| = \prod_{i=1}^{t}(2^{m_i} - 1).$$

Proof. First, we note that $1 = -1$ in R, otherwise -1 is a unit of even order. Consider the group algebra kU, where $k = \mathbb{F}_2$. Since k has characteristic 2 and $|U|$ is odd, Maschke's theorem says that kU is semisimple. There is a ring map $\varphi \colon kU \to R$ carrying every k-linear combination of elements of U to "itself." Now $R' = \mathrm{im}\,\varphi$ is a finite subring of R containing U (for kU is finite); since dropping to a subring cannot create any new

units, we have $U = U(R')$. By Corollary 8.43(iii), the ring R' is semisimple, so that the Wedderburn–Artin Theorem I gives

$$R' \cong \prod_{i=1}^{t} \mathrm{Mat}_{n_i}(\Delta_i),$$

where each Δ_i is a division ring.

Now Δ_i is finite, because R' is finite, and so Δ_i is a finite division ring. By the "other" theorem of Wedderburn, Theorem 8.23, each Δ_i is a field. But $-1 = 1$ in R implies $-1 = 1$ in Δ_i, and so each field Δ_i has characteristic 2; hence,

$$|\Delta_i| = 2^{m_i}$$

for integers $m_i \geq 1$. All the matrix rings must be 1×1, for any matrix ring of larger size must contain an element of order 2, namely, $I + K$, where K has entry 1 in the first position in the bottom row, and all other entries 0. For example,

$$\begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}^2 = \begin{bmatrix} 1 & 0 \\ 2 & 1 \end{bmatrix} = I.$$

Therefore, R' is a direct product of finite fields of characteristic 2, and so $U = U(R')$ is an abelian group whose order is described in the statement. •

It follows, for example, that there is no ring having exactly five units.

The *Jacobson–Chevalley density theorem*, an important generalization of Wedderburn's theorem for certain nonartinian rings, was proved in the 1930s. Call a ring R ***left primitive*** if there exists a faithful simple left R-module S; that is, S is simple and, if $r \in R$ and $rS = \{0\}$, then $r = 0$. It can be proved that commutative primitive rings are fields, while left artinian left primitive rings are simple. Assume now that R is a left primitive ring, that S is a faithful simple left R-module, and that Δ denotes the division ring $\mathrm{End}_R(S)$. The density theorem says that if R is left artinian, then $R \cong \mathrm{Mat}_n(\Delta)$, while if R is not left artinian, then for every integer $n > 0$, there exists a subring R_n of R with $R_n \cong \mathrm{Mat}_n(\Delta)$. We refer the reader to Lam, *A First Course in Noncommutative Rings*, pages 191–193.

The Wedderburn–Artin theorems led to several areas of research, two of which are descriptions of division rings and of finite-dimensional algebras. Division rings will be considered in the context of central simple algebras in Chapter 9 and crossed product algebras in Chapter 10. Let us discuss finite dimensional algebras now.

Thanks to the theorems of Maschke and Molien, the Wedderburn–Artin theorems apply to *ordinary* representations of a finite group G; that is, to kG-modules, where k is a field whose characteristic does not divide $|G|$. We know kG is semisimple in this case. However, *modular* representations, that is, kG-modules for which the characteristic of k does divide $|G|$, arise naturally. For example, if G is a finite p-group, for some prime p, then a minimal normal subgroup N is a vector space over $\mathbb{F}_p$. Now G acts on N (by conjugation), and so N is an $\mathbb{F}_p G$-module. Modular representations are used extensively

in the classification of the finite simple groups. In his study of modular representations, R. Brauer observed that the important modules M are *indecomposable* rather than irreducible. Recall that a module M is indecomposable if there are no nonzero modules A and B with $M = A \oplus B$ (in the ordinary case, a module is indecomposable if and only if it is irreducible [i.e., simple], but this is no longer true in the modular case). When kG is semisimple, Proposition 8.54 says that there are only finitely many indecomposable modules (corresponding to the minimal left ideals). This is not true in the modular case, however. For example, if k is an algebraically closed field of characteristic 2, $k\mathbf{V}$ and kA_4 have infinitely many nonisomorphic indecomposable modules.

A finite-dimensional k-algebra R over a field k is said to have ***finite representation type*** if there are only finitely many nonisomorphic finite-dimensional indecomposable R-modules. D. G. Higman proved that if G is a finite group, then kG has finite representation type for every field k if and only if all its Sylow subgroups G are cyclic. In the 1950s, the following two problems, known as the ***Brauer–Thrall conjectures***, were posed. Let R be a ring not of finite representation type.

(I). Are the dimensions of the indecomposable R-modules unbounded?

(II). Is there a strictly increasing sequence $n_1, n_2, \ldots$ with infinitely many nonisomorphic indecomposable R-modules of dimension n_i for every i?

The positive solution of the first conjecture, by A. V. Roiter in 1968, had a great impact. Shortly thereafter, P. Gabriel introduced graph-theoretic methods, associating finite-dimensional algebras to certain oriented graphs, called *quivers*. He proved that a connected quiver has a finite number of nonisomorphic finite-dimensional representations if and only if the quiver is one of the Dynkin diagrams A_n, D_n, E_6, E_7, or E_8 (*Dynkin diagrams* are multigraphs that describe simple complex Lie algebras; see the discussion on page 778). Gabriel's result can be rephrased in terms of *hereditary* k-algebras A (one-sided ideals are projective A-modules). V. Dlab and C. Ringel extended Gabriel's result to all Dynkin diagrams (of any type A through G). They proved that a finite-dimensional hereditary algebra is of finite representation type if and only if its graph is a finite union of Dynkin diagrams. Moreover, using *Coxeter functors* (which were introduced by I. N. Bernstein, I. M. Gelfand, and V. A. Ponomarev to give a new proof of Gabriel's result), they extended the classification to hereditary algebras of *tame* representation type in terms of the so-called extended Dynkin diagrams (algebras of infinite representation type are divided into those of tame type and those of wild type). A confirmation of the second Brauer–Thrall conjecture for all hereditary algebras followed. A positive solution of Brauer–Thrall II for all (not necessarily hereditary) finite-dimensional algebras over an algebraically closed field follows from the multiplicative basis theorem of R. Bautista, P. Gabriel, A. V. Roiter, and L. Salmerón: Every finite-dimensional k-algebra A of finite representation type has a *multiplicative basis* B: a vector space basis of A such that the product of two basis vectors lies in $B \cup \{0\}$. In fact, they proved that there exist multiplicative bases that contain a complete set of primitive orthogonal idempotents and a basis of each power of the radical. M. Auslander and I. Reiten created a theory involving *almost split sequences* (defined in Chapter 10) and *Auslander–Reiten quivers*. This theory, which generalizes the concept of

Coxeter functors, provides a construction of new indecomposable representations of (arbitrary) finite-dimensional algebras. As of this writing, Auslander–Reiten theory is the most powerful tool in the study of representations of finite-dimensional algebras. For a discussion of these ideas, we refer the reader to Artin–Nesbitt–Thrall, *Rings with Minimum Condition*, Dlab–Ringel, *Indecomposable Representations of Graphs and Algebras*, Memoir AMS #173, 1976, Jacobson, *The Theory of Rings*, Jacobson, *Structure of Rings*, and Drozd-Kirichenko, *Finite Dimensional Algebras*.

EXERCISES

8.36 Let A be an n-dimensional k-algebra over a field k. Prove that A can be imbedded as a k-subalgebra of $\mathrm{Mat}_n(k)$.
Hint. If $a \in A$, define $L_a \colon A \to A$ by $L_a \colon x \mapsto ax$.

8.37 Let G be a finite group, and let k be a commutative ring. Define $\varepsilon \colon kG \to k$ by

$$\varepsilon\Big(\sum_{g \in G} a_g g\Big) = \sum_{g \in G} a_g$$

(this map is called the ***augmentation***, and its kernel, denoted by $\mathcal{G}$, is called the ***augmentation ideal***).

(i) Prove that ε is a kG-map and that $kG/\mathcal{G} \cong k$ as k-algebras. Conclude that $\mathcal{G}$ is a two-sided ideal in kG.

(ii) Prove that $kG/\mathcal{G} \cong V_0(k)$, where $V_0(k)$ is k viewed as a trivial kG-module.
Hint. $\mathcal{G}$ is a two-sided ideal containing $xu - u = (x-1)u \in \mathcal{G}$.

(iii) Use part (ii) to prove that if $kG = \mathcal{G} \oplus V$, then $V = \langle v \rangle$, where $v = a \sum_{g \in G} g$.
Hint. Argue as in Example 8.55.

(iv) Assume that k is a field whose characteristic p does divide $|G|$. Prove that kG is not left semisimple.
Hint. First show that $\varepsilon(v) = 0$, and then show that the short exact sequence

$$0 \to \mathcal{G} \to kG \xrightarrow{\varepsilon} k \to 0$$

does not split.

8.38 If Δ is a division ring, prove that every two minimal left ideals in $\mathrm{Mat}_n(\Delta)$ are isomorphic. (Compare Corollary 8.50.)

8.39 An element a in a ring R is called a ***zero divisor*** if $a \neq 0$ and there exists a nonzero $b \in R$ with $ab = 0$ (more precisely, we call a a left zero divisor and b a right zero divisor). Prove that a left artinian ring R having no zero divisors must be a division ring.

8.40 Let $T \colon V \to V$ be a linear transformation, where V is a vector space over a field k, and let $k[T]$ be defined by

$$k[T] = k[x]/(m(x)),$$

where $m(x)$ is the minimum polynomial of T.

(i) If $m(x) = \prod_p p(x)^{e_p}$, where the $p(x) \in k[x]$ are distinct irreducible polynomials and $e_p \geq 1$, prove that $k[T] \cong \prod_p k[x]/(p(x)^{e_p})$.

(ii) Prove that $k[T]$ is a semisimple ring if and only if $m(x)$ is a product of distinct linear factors. (In linear algebra, we show that this last condition is equivalent to T being *diagonalizable*; that is, any matrix of T [arising from some choice of basis of T] is similar to a diagonal matrix.)

8.41 Find $\mathbb{C}G$ if $G = D_8$, the dihedral group of order 8.

8.42 Find $\mathbb{C}G$ if $G = A_4$.
Hint. A_4 has four conjugacy classes.

8.43 (i) Let k be a field, and view sgn: $S_n \to \{\pm 1\} \leq k$. Define $\mathrm{Sig}(k)$ to be k made into a kS_n-module (as in Proposition 8.37): If $\gamma \in S_n$ and $a \in k$, then $\gamma a = \mathrm{sgn}(\gamma)a$. Prove that $\mathrm{Sig}(k)$ is an irreducible kS_n-module, and if k does not have characteristic 2, then $\mathrm{Sig}(k) \ncong V_0(k)$.

(ii) Find $\mathbb{C}S_5$.
Hint. There are five conjugacy classes in S_5.

8.44 Let G be a finite group, and let k and K be algebraically closed fields whose characteristics p and q, respectively, do not divide $|G|$.

(i) Prove that kG and KG have the same number of simple components.

(ii) Prove that the degrees of the irreducible representations of G over k are the same as the degrees of the irreducible representations of G over K.

8.4 TENSOR PRODUCTS

We now introduce a new notion,[7] tensor products, that is used to construct *induced representations* (which extend representations of subgroups to representations of the whole group). Tensor products are also useful in other areas of algebra as well; for example, they are involved in bilinear forms, the adjoint isomorphism, free algebras, exterior algebra, and determinants. The reader who wishes to see the impact of the Wedderburn–Artin and Maschke theorems on groups without this interruption can proceed directly to the next section, for the first application we shall give—Burnside's theorem—does not use induced representations in its proof. On the other hand, we shall also prove a theorem of Frobenius that does use induced representations.

If k is a field and H be a subgroup of a group G, then a k-representation of H is the same thing as a kH-module, and a k-representation of G is the same thing as a kG-module. If we could force a kH-module M to be a kG-module, then we would be able to create a representation of the big group G from a representation of a subgroup. More generally, if A is a subring of a ring R, we may want to force an A-module M to be an R-module. If M is generated as an A-module by a set X, then each $m \in M$ has an expression of the form $m = \sum_i a_i x_i$, where $a_i \in A$ and $x_i \in X$. Perhaps we could create an R-module containing M by taking all expressions of the form $\sum_i r_i x_i$ for $r_i \in R$. This naive approach is doomed

[7] Tensor products of R-modules, where R is commutative, could have been presented in Chapter 7. However, I believe that the best exposition delays the introduction of noncommutative rings to the present chapter. Consequently, putting tensor products earlier would have forced me to construct them in two stages: first over commutative rings in Chapter 7, then over general rings now. This is not a good idea.

to failure. For example, a cyclic group $G = \langle g \rangle$ of finite order n is a $\mathbb{Z}$-module; can we make it into a $\mathbb{Q}$-module? A $\mathbb{Q}$-module V is a vector space over $\mathbb{Q}$, and it is easy to see that if $v \in V$ and $q \in \mathbb{Q}$, then $qv = 0$ if and only if $q = 0$ or $v = 0$. If we could create a rational vector space V containing G in the naive way described in the previous paragraph, then $ng = 0$ would imply $g = 0$ in V! Our goal of adjoining scalars to obtain a module over a larger ring still has merit but, plainly, we cannot be so cavalier about its construction. The proper way to deal with such matters is with *tensor products*.

One of the most compelling reasons to introduce tensor products comes from algebraic topology, where we assign to every topological space X a sequence of *homology groups* $H_n(X)$ for all $n \geq 0$ that are of basic importance. The *Künneth formula* computes the homology groups of the cartesian product $X \times Y$ of two topological spaces in terms of the tensor product of the homology groups of the factors X and Y.

Definition. Let R be a ring, let A_R be a right R-module, let ${}_RB$ be a left R-module, and let G be an (additive) abelian group. A function $f\colon A \times B \to G$ is called ***R-biadditive*** if, for all $a, a' \in A$, $b, b' \in B$, and $r \in R$, we have

$$f(a + a', b) = f(a, b) + f(a', b);$$
$$f(a, b + b') = f(a, b) + f(a, b');$$
$$f(ar, b) = f(a, rb).$$

An R-biadditive function is also called a ***pairing***.

If R is *commutative* and A, B, and M are R-modules, then a function $f\colon A \times B \to M$ is called ***R-bilinear*** if f is R-biadditive and also

$$f(ar, b) = f(a, rb) = rf(a, b).$$

Example 8.73.

(i) If R is a ring, then its multiplication $\mu\colon R \times R \to R$ is R-biadditive; the first two axioms are the right and left distributive laws, while the third axiom is associativity:

$$\mu(ar, b) = (ar)b = a(rb) = \mu(a, rb).$$

If R is a commutative ring, then μ is R-bilinear, for $(ar)b = a(rb) = r(ab)$.

(ii) If ${}_RM$ is a left R-module, then its scalar multiplication $\sigma\colon R \times M \to M$ is R-biadditive; if R is a commutative ring, then σ is R-bilinear.

(iii) If M_R and N_R are right R-modules, then $\mathrm{Hom}_R(M, N)$ is a left R-module if, for $f \in \mathrm{Hom}_R(M, N)$ and $r \in R$, we define $rf\colon M \to N$ by

$$rf\colon m \mapsto f(mr).$$

The reader may show that this does make Hom into a left R-module; moreover, we can now see that *evaluation* $e\colon M \times \mathrm{Hom}_R(M, N) \to N$, given by $(m, f) \mapsto f(m)$, is R-biadditive.

The dual space V^* of a vector space V over a field k gives a special case of this construction: Evaluation $V \times V^* \to k$ is R-bilinear.

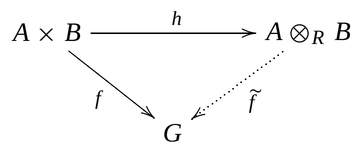
(iv) If $G^* = \mathrm{Hom}_{\mathbb{Z}}(G, \mathbb{Q}/\mathbb{Z})$ is the Pontrjagin dual of an abelian group G, then evaluation $G \times G^* \to \mathbb{Q}/\mathbb{Z}$ is $\mathbb{Z}$-bilinear. ◀

Tensor products convert biadditive functions into linear ones.

Definition. Given a ring R and modules A_R and ${}_RB$, then their ***tensor product*** is an abelian group $A \otimes_R B$ and an R-biadditive function

$$h \colon A \times B \to A \otimes_R B$$

such that, for every abelian group G and every R-biadditive $f \colon A \times B \to G$, there exists a unique $\mathbb{Z}$-homomorphism $\widetilde{f} \colon A \otimes_R B \to G$ making the following diagram commute.



If a tensor product of A and B exists, then it is unique to isomorphism, for it has been defined as a solution to a universal mapping problem (see the proof of Proposition 7.27 on page 448).

Quite often, we denote $A \otimes_R B$ by $A \otimes B$ when $R = \mathbb{Z}$.

Proposition 8.74. *If R is a ring and A_R and ${}_RB$ are modules, then their tensor product exists.*

Proof. Let F be the free abelian group with basis $A \times B$; that is, F is free on all ordered pairs (a, b), where $a \in A$ and $b \in B$. Define S to be the subgroup of F generated by all elements of the following types:

$$\begin{aligned}(a, b + b') - (a, b) - (a, b');\\ (a + a', b) - (a, b) - (a', b);\\ (ar, b) - (a, rb).\end{aligned}$$

Define $A \otimes_R B = F/S$, denote the coset $(a, b) + S$ by $a \otimes b$, and define

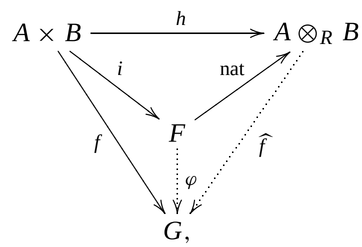
$$h \colon A \times B \to A \otimes_R B \quad \text{by} \quad h \colon (a, b) \mapsto a \otimes b$$

(thus, h is the restriction of the natural map $F \to F/S$). We have the following identities in $A \otimes_R B$:

$$\begin{aligned}a \otimes (b + b') &= a \otimes b + a \otimes b';\\ (a + a') \otimes b &= a \otimes b + a' \otimes b;\\ ar \otimes b &= a \otimes rb.\end{aligned}$$

It is now obvious that h is R-biadditive.

Consider the following diagram, where G is an abelian group and f is R-biadditive:



where $i\colon A \times B \to F$ is the inclusion. Since F is free abelian with basis $A \times B$, there exists a homomorphism $\varphi\colon F \to G$ with $\varphi(a, b) = f(a, b)$ for all (a, b); now $S \subseteq \ker \varphi$ because f is R-biadditive, and so φ induces a map $\widehat{f}\colon A \otimes_R B \to G$ (because $A \otimes_R B = F/S$) by

$$\widehat{f}(a \otimes b) = \widehat{f}((a, b) + S) = \varphi(a, b) = f(a, b).$$

This equation may be rewritten as $\widehat{f}h = f$; that is, the diagram commutes. Finally, $\widehat{f}$ is unique because $A \otimes_R B$ is generated by the set of all $a \otimes b$'s. •

Remark. Since $A \otimes_R B$ is generated by the elements of the form $a \otimes b$, every $u \in A \otimes_R B$ has the form

$$u = \sum_i a_i \otimes b_i .$$

This expression for u is not unique; for example, there are expressions

$$\begin{aligned} 0 &= a \otimes (b + b') - a \otimes b - a \otimes b' \\ &= (a + a') \otimes b - a \otimes b - a' \otimes b \\ &= ar \otimes b - a \otimes rb. \end{aligned}$$

Therefore, given some abelian group G, we must be suspicious of a *definition* of a map $g\colon A \otimes_R B \to G$ that is given by specifying g on the generators $a \otimes b$; such a "function" g may not be well-defined because elements have many expressions in terms of these generators. In essence, g is only defined on F (the free abelian group with basis $A \times B$), and we must still show that $g(S) = \{0\}$, because $A \otimes_R B = F/S$. The simplest (and safest!) procedure is to define an R-biadditive function on $A \times B$, and it will yield a (well-defined) homomorphism. We illustrate this procedure in the next proof. ◂

Proposition 8.75. *Let $f\colon A_R \to A'_R$ and $g\colon {}_RB \to {}_RB'$ be maps of right R-modules and left R-modules, respectively. Then there is a unique $\mathbb{Z}$-homomorphism, denoted by $f \otimes g\colon A \otimes_R B \to A' \otimes_R B'$, with*

$$f \otimes g\colon a \otimes b \mapsto f(a) \otimes g(b).$$

Proof. The function $\varphi\colon A \times B \to A' \otimes_R B'$, given by $(a, b) \mapsto f(a) \otimes g(b)$, is easily seen to be an R-biadditive function. For example,

$$\varphi\colon (ar, b) \mapsto f(ar) \otimes g(b) = f(a)r \otimes g(b)$$

and

$$\varphi\colon (a, rb) \mapsto f(a) \otimes g(rb) = f(a) \otimes rg(b);$$

these are equal because of the identity $a'r \otimes b' = a' \otimes rb'$ in $A' \otimes_R B'$. The biadditive function φ yields a unique homomorphism $A \otimes_R B \to A' \otimes_R B'$ taking

$$a \otimes b \mapsto f(a) \otimes g(b). \quad \bullet$$

Corollary 8.76. *Given maps of right R-modules, $A \xrightarrow{f} A' \xrightarrow{f'} A''$, and maps of left R-modules, $B \xrightarrow{g} B' \xrightarrow{g'} B''$,*

$$(f' \otimes g')(f \otimes g) = f'f \otimes g'g.$$

Proof. Both maps take $a \otimes b \mapsto f'f(a) \otimes g'g(b)$, and so the uniqueness of such a homomorphism gives the desired equation. $\bullet$

Theorem 8.77. *Given A_R, there is an additive functor $F_A\colon {}_R\mathbf{Mod} \to \mathbf{Ab}$, defined by*

$$F_A(B) = A \otimes_R B \quad \textit{and} \quad F_A(g) = 1_A \otimes g,$$

where $g\colon B \to B'$ is a map of left R-modules.

Proof. First, note that F_A preserves identities: $F_A(1_B) = 1_A \otimes 1_B$ is the identity $1_{A \otimes B}$, because it fixes every generator $a \otimes b$. Second, F_A preserves composition:

$$F_A(g'g) = 1_A \otimes g'g = (1_A \otimes g')(1_A \otimes g) = F_A(g')F_A(g),$$

by Corollary 8.76. Therefore, F_A is a functor.

To see that F_A is additive, we must show that $F_A(g + h) = F_A(g) + F_A(h)$, where $g, h\colon B \to B'$; that is, $1_A \otimes (g + h) = 1_A \otimes g + 1_A \otimes h$. This is also easy, for both these maps send $a \otimes b \mapsto a \otimes g(b) + a \otimes h(b)$. $\bullet$

We denote the functor F_A by $A \otimes_R \;\;$. Of course, there is a similar result if we fix a left R-module B: There is an additive functor $\;\; \otimes_R B\colon \mathbf{Mod}_R \to \mathbf{Ab}$.

Corollary 8.78. *If $f\colon M \to M'$ and $g\colon N \to N'$ are, respectively, isomorphisms of right and left R-modules, then $f \otimes g\colon M \otimes_R N \to M' \otimes_R N'$ is an isomorphism of abelian groups.*

Proof. Now $f \otimes 1_{N'}$ is the value of the functor $F_{N'}$ on the isomorphism f, and hence $f \otimes 1_{N'}$ is an isomorphism; similarly, $1_M \otimes g$ is an isomorphism. By Corollary 8.76, we have $f \otimes g = (f \otimes 1_{N'})(1_M \otimes g)$. Therefore, $f \otimes g$ is an isomorphism, being the composite of isomorphisms. •

Before continuing with properties of tensor products, we pause to discuss a technical point. In general, the tensor product of two modules is only an abelian group; is it ever a module? If so, do the tensor product functors then take values in a module category, not merely in **Ab**? That is, is $1 \otimes f$ always a map of modules?

Definition. Let R and S be rings and let M be an abelian group. Then M is an (R, S)-***bimodule***, denoted by ${}_R M_S$, if M is a left R-module and a right S-module, and the two scalar multiplications are related by an associative law:

$$r(ms) = (rm)s$$

for all $r \in R$, $m \in M$, and $s \in S$.

If M is an (R, S)-bimodule, it is permissible to write rms with no parentheses, for the definition of bimodule says that the two possible associations agree.

Example 8.79.

(i) Every ring R is an (R, R)-bimodule; the extra identity is just the associativity of multiplication in R.

(ii) Every two-sided ideal in a ring R is an (R, R)-bimodule.

(iii) If M is a left R-module (i.e., if $M = {}_R M$), then M is an $(R, \mathbb{Z})$-bimodule; that is, $M = {}_R M_{\mathbb{Z}}$. Similarly, a right R-module N is a bimodule ${}_{\mathbb{Z}} N_R$.

(iv) If R is commutative, then every left (or right) R-module is an (R, R)-bimodule. In more detail, if $M = {}_R M$, define a new scalar multiplication $M \times R \to M$ by $(m, r) \mapsto rm$. To see that M is a right R-module, we must show that $m(rr') = (mr)r'$, that is, $(rr')m = r'(rm)$, and this is so because $rr' = r'r$. Finally, M is an (R, R)-bimodule because both $r(mr')$ and $(rm)r'$ are equal to $(rr')m$.

(v) In Example 8.6, we made any left kG-module M into a right kG-module by defining $mg = g^{-1}m$ for every $m \in M$ and every g in the group G. Even though M is both a left and right kG-module, it is usually not a (kG, kG)-bimodule because the required associativity formula may not hold. In more detail, let $g, h \in G$ and let $m \in M$. Now $g(mh) = g(h^{-1}m) = (gh^{-1})m$; on the other hand, $(gm)h = h^{-1}(gm) = (h^{-1}g)m$. To see that these can be different, take $M = kG$, $m = 1$, and g and h noncommuting elements of G. ◂

The next lemma solves the problem of extending scalars.

Lemma 8.80. *Given a bimodule ${}_S A_R$ and a left module ${}_R B$, then the tensor product $A \otimes_R B$ is a left S-module, where*

$$s(a \otimes b) = (sa) \otimes b.$$

Similarly, given A_R and ${}_R B_S$, the tensor product $A \otimes_R B$ is a right S-module, where $(a \otimes b)s = a \otimes (bs)$.

In particular, if k is a commutative ring and A is a k-algebra, then $A \otimes_k B$ is a left A-module.

Proof. For fixed $s \in S$, the multiplication $\mu_s \colon A \to A$, defined by $a \mapsto sa$, is an R-map, for A being a bimodule gives

$$\mu_s(ar) = s(ar) = (sa)r = \mu_s(a)r.$$

If $F = \quad \otimes_R B \colon \mathbf{Mod}_R \to \mathbf{Ab}$, then $F(\mu_s) \colon A \otimes_R B \to A \otimes_R B$ is a (well-defined) $\mathbb{Z}$-homomorphism. Thus, $F(\mu_s) = \mu_s \otimes 1_B \colon a \otimes b \mapsto (sa) \otimes b$, and so the formula in the statement of the lemma makes sense. It is now straightforward to check that the module axioms do hold for $A \otimes_R B$.

The last statement follows because a k-algebra A is an (A, k)-bimodule. •

For example, if V and W are vector spaces over a field k, then their tensor product $V \otimes_k W$ is also a vector space over k.

After a while, we see that proving properties of tensor products is just a matter of showing that the obvious maps are, indeed, well-defined functions.

We have made some progress in our original problem: Given a left k-module M, where k is a subring of a ring K, we can create a left K-module from M by *extending scalars*; that is, Lemma 8.80 shows that $K \otimes_k M$ is a left K-module, for K is a (K, k)-bimodule. However, we must still investigate, among other things, why a left k-module M may not be imbedded in $K \otimes_k M$, where k is a subring of a ring K.

The following special case of extending scalars is important for representations. If H is a subgroup of a group G and if $\rho \colon H \to \mathrm{GL}(V)$ is a k-representation, then $\rho \colon H \to \mathrm{GL}(V)$ equips V with a left kH-module structure. We call $V^G = kG \otimes_{kH} V$ the *induced module*. Note that kG is a right kH-module (it is even a right kG-module), and so the tensor product $V^G = kG \otimes_{kH} V$ makes sense; moreover, V^G is a left kG-module, by Lemma 8.80. We will investigate this construction more carefully later in this chapter (see *induced modules* on page 624).

Corollary 8.81.

(i) *Given a bimodule ${}_S A_R$, then the functor $F_A = A \otimes_R \quad \colon {}_R\mathbf{Mod} \to \mathbf{Ab}$ actually takes values in ${}_S\mathbf{Mod}$.*

(ii) *If R is a commutative ring, then $A \otimes_R B$ is an R-module, where*

$$r(a \otimes b) = (ra) \otimes b = a \otimes rb$$

for all $r \in R$, $a \in A$, and $b \in B$.

(iii) *If R is a commutative ring, $r \in R$, and $\mu_r \colon B \to B$ is multiplication by r, then*

$$1_A \otimes \mu_r \colon A \otimes_R B \to A \otimes_R B$$

is also multiplication by r.

Proof. (i) By the lemma, we know that $A \otimes_R B$ is a left S-module, where $s(a \otimes b) = (sa) \otimes b$, and so it suffices to show that if $g\colon B \to B'$ is a map of left R-modules, then $F_A(g) = 1_A \otimes g$ is an S-map. But

$$\begin{aligned}(1_A \otimes g)[s(a \otimes b)] &= (1_A \otimes g)[(sa) \otimes b]\\ &= (sa) \otimes gb\\ &= s(a \otimes gb) \qquad\qquad \text{by Lemma 8.80}\\ &= s(1_A \otimes g)(a \otimes b).\end{aligned}$$

(ii) Since R is commutative, we may regard A as an (R, R)-bimodule by defining $ar = ra$. Lemma 8.80 now gives

$$r(a \otimes b) = (ra) \otimes b = (ar) \otimes b = a \otimes rb.$$

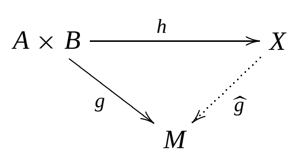
(iii) This statement merely sees the last equation $a \otimes rb = r(a \otimes b)$ from a different viewpoint:

$$(1_A \otimes \mu_r)(a \otimes b) = a \otimes rb = r(a \otimes b). \quad \bullet$$

We have defined R-biadditive functions for arbitrary, possibly noncommutative, rings R, whereas we have defined R-bilinear functions only for commutative rings. Tensor product was defined as the solution of a certain universal mapping problem involving R-biadditive functions; we now consider the analogous problem for R-bilinear functions when R is commutative.

Here is a provisional definition, soon to be seen unnecessary.

Definition. If k is a commutative ring, then a ***k-bilinear product*** is a k-module X and a k-bilinear function $h\colon A \times B \to X$ such that, for every k-module M and every k-bilinear function $g\colon A \times B \to M$, there exists a unique k-homomorphism $\widehat{g}\colon X \to M$ making the following diagram commute.



The next result shows that k-bilinear products exist, but that they are nothing new.

Proposition 8.82. *If k is a commutative ring and A and B are k-modules, then the k-module $A \otimes_k B$ is a k-bilinear product.*

Proof. We show that $X = A \otimes_k B$ provides the solution if we define $h(a, b) = a \otimes b$; note that h is also k-bilinear, thanks to Corollary 8.81. Since g is k-bilinear, it is k-biadditive, and so there does exist a $\mathbb{Z}$-homomorphism $\widehat{g}\colon A \otimes_k B \to M$ with $\widehat{g}(a \otimes b) = g(a, b)$ for all $(a, b) \in A \times B$. We need only show that $\widehat{g}$ is a k-map. If $u \in k$,

$$\begin{aligned} \widehat{g}(u(a \otimes b)) &= \widehat{g}((ua) \otimes b) \\ &= g(ua, b) \\ &= ug(a, b) && \text{for } g \text{ is } k\text{-bilinear} \\ &= u\widehat{g}(a \otimes b). \quad \bullet \end{aligned}$$

As a consequence of the proposition, the term *bilinear product* is unnecessary, and we shall call it the *tensor product* instead.

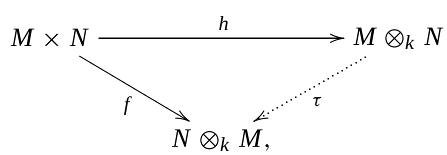
In contrast to the Hom functors, the tensor functors obey certain commutativity and associativity laws.

Proposition 8.83 (Commutativity). *If k is a commutative ring and M and N are k-modules, then there is a k-isomorphism*

$$\tau\colon M \otimes_k N \to N \otimes_k M$$

with $\tau : m \otimes n \mapsto n \otimes m$.

Proof. First, Corollary 8.81 shows that both $M \otimes_k N$ and $N \otimes_k M$ are k-modules. Consider the diagram



where $f(m, n) = n \otimes m$. It is easy to see that f is k-bilinear, and so there is a unique k-map $\tau : M \otimes_k N \to N \otimes_k M$ with $\tau : m \otimes n \mapsto n \otimes m$. A similar diagram, interchanging the roles of $M \otimes_k N$ and $N \otimes_k M$, gives a k-map in the reverse direction taking $n \otimes m \mapsto m \otimes n$. Both composites of these maps are obviously identity maps, and so τ is a k-isomorphism. $\bullet$

Proposition 8.84 (Associativity). *Given A_R, ${}_R B_S$, and ${}_S C$, there is an isomorphism*

$$\theta\colon A \otimes_R (B \otimes_S C) \cong (A \otimes_R B) \otimes_S C$$

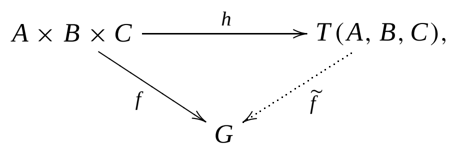
given by

$$a \otimes (b \otimes c) \mapsto (a \otimes b) \otimes c.$$

Proof. Define a ***triadditive*** function $f : A \times B \times C \to G$, where G is an abelian group, to be a function that is additive in each of the three variables (when we fix the other two),

$$f(ar, b, c) = f(a, rb, c), \quad \text{and} \quad f(a, bs, c) = f(a, b, sc),$$

for all $r \in R$ and $s \in S$. Consider the univeral mapping problem described by the diagram



where G is an abelian group, f is triadditive, and $\widetilde{f}$ is a $\mathbb{Z}$-homomorphism. As for biadditive functions and tensor products of two modules, define $T(A, B, C) = F/N$, where F is the free abelian group on all ordered triples $(a, b, c) \in A \times B \times C$, and N is the obvious subgroup of relations. Define $h\colon A \times B \times C \to T(A, B, C)$ by

$$h\colon (a, b, c) \mapsto (a, b, c) + N$$

(denote $(a, b, c) + N$ by $a \otimes b \otimes c$). A routine check shows that this construction does give a solution to the universal mapping problem for triadditive functions.

We now show that $A \otimes_R (B \otimes_S C)$ is another solution to this universal problem. Define a triadditive function $\eta\colon A \times B \times C \to A \otimes_R (B \otimes_S C)$ by $\eta\colon (a, b, c) \mapsto a \otimes (b \otimes c)$; we must find a homomorphism $\widetilde{f}\colon A \otimes_R (B \otimes_S C) \to G$ with $\widetilde{f}\eta = f$. For each $a \in A$, the S-biadditive function $f_a\colon B \times C \to G$, defined by $(b, c) \mapsto f(a, b, c)$, gives a unique homomorphism $\widetilde{f_a}\colon B \otimes_S C \to G$ taking $b \otimes c \mapsto f(a, b, c)$. If $a, a' \in A$, then $\widetilde{f}_{a+a'}(b \otimes c) = f(a + a', b, c) = f(a, b, c) + f(a', b, c) = \widetilde{f_a}(b \otimes c) + \widetilde{f_{a'}}(b \otimes c)$. It follows that the function $\varphi\colon A \times (B \otimes_S C) \to G$, defined by $\varphi(a, b \otimes c) = \widetilde{f_a}(b \otimes c)$, is additive in both variables. It is R-biadditive, for if $r \in R$, then $\varphi(ar, b \otimes c) = \widetilde{f_{ar}}(b \otimes c) = f(ar, b, c) = f(a, rb, c) = \widetilde{f_a}(rb \otimes c) = \varphi(a, r(b \otimes c))$. Therefore, there is a unique homomorphism $\widetilde{f}\colon A \otimes_R (B \otimes_S C) \to G$ with $a \otimes (b \otimes c) \mapsto \varphi(a, b \otimes c) = f(a, b, c)$; that is, $\widetilde{f}\eta = f$. Uniqueness of solutions to universal mapping problems shows that there is an isomorphism $T(A, B, C) \to A \otimes_R (B \otimes_S C)$ with $a \otimes b \otimes c \mapsto a \otimes (b \otimes c)$. Similarly, $T(A, B, C) \cong (A \otimes_R B) \otimes_S C$ via $a \otimes b \otimes c \mapsto (a \otimes b) \otimes c$, and so $A \otimes_R (B \otimes_S C) \cong (A \otimes_R B) \otimes_S C$ via $a \otimes (b \otimes c) \mapsto (a \otimes b) \otimes c$. •

Remark. That the elements $a \otimes b \otimes c \in T(A, B, C)$ have no parentheses will be exploited in the next chapter when we construct tensor algebras. ◂

We now present properties of tensor products that will help us compute them. First, we give a result about Hom, and then we give the analogous result for tensor.

Recall Exercise 8.34 on page 549: For any left R-module M, for any $f \in \text{Hom}_R(R, M)$, and for any $r, s \in R$, define

$$rf\colon s \mapsto f(sr).$$

Using the fact that a ring R is an (R, R)-bimodule, we can check that rf is an R-map and that $\text{Hom}_R(R, M)$ is a left R-module. We incorporate this into the next result.

Proposition 8.85. *If M is a left R-module, then $\operatorname{Hom}_R(R, M)$ is a left R-module, and there is an R-isomorphism $\varphi_M \colon \operatorname{Hom}_R(R, M) \to M$, given by $\varphi_M(f) = f(1)$. Indeed, $\varphi = \{\varphi_M\}$ is a natural equivalence between $\operatorname{Hom}_R(R, \quad)$ and the identity functor on ${}_R\mathbf{Mod}$.*

Proof. Adapt the proof of Proposition 7.102. •

Proposition 8.86. *For every left R-module M, there is an R-isomorphism*

$$\theta_M \colon R \otimes_R M \to M$$

with $\theta_M \colon r \otimes m \mapsto rm$. Indeed, $\theta = \{\theta_M\}$ is a natural equivalence between $R \otimes_R \quad$ and the identity functor on ${}_R\mathbf{Mod}$.

Proof. The function $R \times M \to M$, given by $(r, m) \mapsto rm$, is R-biadditive, and so there is an R-homomorphism $\theta \colon R \otimes_R M \to M$ with $r \otimes m \mapsto rm$ [we are using the fact that R is an (R, R)-bimodule]. To see that θ is an R-isomorphism, it suffices to find a $\mathbb{Z}$-homomorphism $f \colon M \to R \otimes_R M$ with θf and $f\theta$ identity maps (for it is now only a question of whether the *function* θ is a bijection). Such a $\mathbb{Z}$-map is given by $f \colon m \mapsto 1 \otimes m$.

To see that the isomorphisms θ_M constitute a natural equivalence, we must show, for any module homomorphism $h \colon M \to N$, that the following diagram commutes.

$$\begin{array}{ccc} R \otimes_R M & \xrightarrow{1 \otimes h} & R \otimes_R N \\ {\scriptstyle \theta_M}\downarrow & & \downarrow{\scriptstyle \theta_N} \\ M & \xrightarrow{\quad h \quad} & N \end{array}$$

It suffices to look at a generator $r \otimes m$ of $R \otimes_R M$. Going clockwise, $r \otimes m \mapsto r \otimes h(m) \mapsto rh(m)$, while going counterclockwise, $r \otimes m \mapsto rm \mapsto h(rm)$. These agree, for h is an R-map, so that $h(rm) = rh(m)$. •

The next theorem says that tensor product preserves arbitrary direct sums.

Theorem 8.87. *Given a right module A_R and left R-modules $\{{}_R B_i : i \in I\}$, there is a $\mathbb{Z}$-isomorphism*

$$\varphi \colon A \otimes_R \sum_{i \in I} B_i \to \sum_{i \in I} (A \otimes_R B_i)$$

with $\varphi \colon a \otimes (b_i) \mapsto (a \otimes b_i)$. Moreover, if R is commutative, then φ is an R-isomorphism.

Proof. Since the function $f \colon A \times \big(\sum_i B_i\big) \to \sum_i (A \otimes_R B_i)$, given by $f \colon (a, (b_i)) \mapsto (a \otimes b_i)$ is R-biadditive, there exists a $\mathbb{Z}$-homomorphism

$$\varphi \colon A \otimes_R \Big(\sum_i B_i\Big) \to \sum_i (A \otimes_R B_i)$$

with $\varphi\colon a \otimes (b_i) \mapsto (a \otimes b_i)$. If R is commutative, then $A \otimes_R \left(\sum_{i \in I} B_i\right)$ and $\sum_{i \in I}\left(A \otimes_R B_i\right)$ are R-modules and φ is an R-map (for φ is the function given by the universal mapping problem in Proposition 8.82).

To see that φ is an isomorphism, we give its inverse. Denote the injection $B_j \to \sum_i B_i$ by λ_j [where $\lambda_j(b_j) \in \sum_i B_i$ has jth coordinate b_j and all other coordinates 0], so that $1_A \otimes \lambda_j\colon A \otimes_R B_j \to A \otimes_R \left(\sum_i B_i\right)$. That direct sum is the coproduct in ${}_R\mathbf{Mod}$ gives a homomorphism $\theta\colon \sum_i (A \otimes_R B_i) \to A \otimes_R \left(\sum_i B_i\right)$ with $\theta\colon (a \otimes b_i) \mapsto a \otimes \sum_i \lambda_i(b_i)$. It is now routine to check that θ is the inverse of φ, so that φ is an isomorphism. •

There is a theorem of C. E. Watts (see Rotman, *An Introduction to Homological Algebra*, page 77) saying that if $T\colon {}_R\mathbf{Mod} \to \mathbf{Ab}$ is a (covariant) right exact functor that preserves direct sums, then there is a right R-module A so that F is naturally equivalent to $A \otimes_R \;$.

Example 8.88.
Let k be a field and let V and W be k-modules; that is, V and W are vector spaces over k. Now W is a free k-module; say, $W = \sum_{i \in I} \langle w_i \rangle$, where $\{w_i : i \in I\}$ is a basis of W. Therefore, $V \otimes_k W \cong \sum_{i \in I} V \otimes_k \langle w_i \rangle$. Similarly, $V = \sum_{j \in J} \langle v_j \rangle$, where $\{v_j : j \in J\}$ is a basis of V and, for each i, $V \otimes_k \langle w_i \rangle \cong \sum_{j \in J} \langle v_j \rangle \otimes_k \langle w_i \rangle$. But the one-dimensional vector spaces $\langle v_j \rangle$ and $\langle w_i \rangle$ are isomorphic to k, and Proposition 8.86 gives $\langle v_j \rangle \otimes_k \langle w_i \rangle \cong \langle v_j \otimes w_i \rangle$. Hence, $V \otimes_k W$ is a vector space over k having $\{v_j \otimes w_i : i \in I \text{ and } j \in J\}$ as a basis. In case both V and W are finite-dimensional, we have

$$\dim(V \otimes_k W) = \dim(V)\dim(W). \quad ◂$$

Example 8.89.
We now show that there may exist elements in a tensor product $V \otimes_k V$ that cannot be written in the form $u \otimes w$ for $u, w \in V$.

Let v_1, v_2 be a basis of a two-dimensional vector space V over a field k. As in Example 8.88, a basis for $V \otimes_k V$ is

$$v_1 \otimes v_1,\ v_1 \otimes v_2,\ v_2 \otimes v_1,\ v_2 \otimes v_2.$$

We claim that there do not exist $u, w \in V$ with $v_1 \otimes v_2 + v_2 \otimes v_1 = u \otimes w$. Otherwise, write u and w in terms of v_1 and v_2:

$$\begin{aligned} v_1 \otimes v_2 + v_2 \otimes v_1 &= u \otimes w \\ &= (av_1 + bv_2) \otimes (cv_1 + dv_2) \\ &= acv_1 \otimes v_1 + adv_1 \otimes v_2 + bcv_2 \otimes v_1 + bdv_2 \otimes v_2. \end{aligned}$$

By linear independence of the basis,

$$ac = 0 = bd \quad \text{and} \quad ad = 1 = bc.$$

The first equation gives $a = 0$ or $c = 0$, and either possibility, when substituted into the second equation, gives $0 = 1$. ◂

As a consequence of Theorem 8.87, if

$$0 \to B' \xrightarrow{i} B \xrightarrow{p} B'' \to 0$$

is a split short exact sequence of left R-modules, then, for every right R-module A,

$$0 \to A \otimes_R B' \xrightarrow{1_A \otimes i} A \otimes_R B \xrightarrow{1_A \otimes p} A \otimes_R B'' \to 0$$

is also a split short exact sequence. What if the exact sequence is not split?

Theorem 8.90 (Right Exactness). *Let A be a right R-module, and let*

$$B' \xrightarrow{i} B \xrightarrow{p} B'' \to 0$$

be an exact sequence of left R-modules. Then

$$A \otimes_R B' \xrightarrow{1_A \otimes i} A \otimes_R B \xrightarrow{1_A \otimes p} A \otimes_R B'' \to 0$$

is an exact sequence of abelian groups.

Remark.

(i) The absence of $0 \to$ at the beginning of the sequence will be discussed later; clearly this has something to do with our initial problem of imbedding a group G in a vector space over $\mathbb{Q}$.

(ii) We will give a nicer proof of this theorem once we prove the adjoint isomorphism (see Proposition 8.100) ◂

Proof. There are three things to check.
(i) $\text{im}(1 \otimes i) \subseteq \ker(1 \otimes p)$.
It suffices to prove that the composite is 0; but

$$(1 \otimes p)(1 \otimes i) = 1 \otimes pi = 1 \otimes 0 = 0.$$

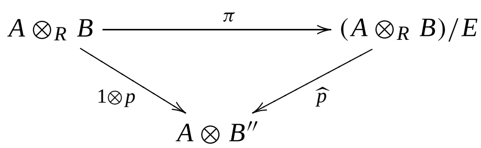
(ii) $\ker(1 \otimes p) \subseteq \text{im}(1 \otimes i)$.
Let $E = \text{im}(1 \otimes i)$. By part (i), $E \subseteq \ker(1 \otimes p)$, and so $1 \otimes p$ induces a map $\widehat{p} \colon (A \otimes B)/E \to A \otimes B''$ with

$$\widehat{p} \colon a \otimes b + E \mapsto a \otimes pb,$$

where $a \in A$ and $b \in B$. Now if $\pi : A \otimes B \to (A \otimes B)/E$ is the natural map, then

$$\widehat{p}\pi = 1 \otimes p,$$

for both send $a \otimes b \mapsto a \otimes pb$.



Suppose we show that $\widehat{p}$ is an isomorphism. Then

$$\ker(1 \otimes p) = \ker \widehat{p}\pi = \ker \pi = E = \operatorname{im}(1 \otimes i),$$

and we are done. To see that $\widehat{p}$ is, indeed, an isomorphism, we construct its inverse $A \otimes B'' \to (A \otimes B)/E$. Define

$$f \colon A \times B'' \to (A \otimes B)/E$$

as follows. If $b'' \in B''$, there is $b \in B$ with $pb = b''$, because p is surjective; let

$$f \colon (a, b'') \mapsto a \otimes b.$$

Now f is well-defined: If $pb_1 = b''$, then $p(b - b_1) = 0$ and $b - b_1 \in \ker p = \operatorname{im} i$. Thus, there is $b' \in B'$ with $ib' = b - b_1$, and hence $a \otimes (b - b_1) = a \otimes ib' \in \operatorname{im}(1 \otimes i) = E$. Clearly, f is R-biadditive, and so the definition of tensor product gives a homomorphism $\widehat{f} \colon A \otimes B'' \to (A \otimes B)/E$ with $\widehat{f}(a \otimes b'') = a \otimes b + E$. The reader may check that $\widehat{f}$ is the inverse of $\widehat{p}$, as desired.

(iii) $1 \otimes p$ is surjective.

If $\sum a_i \otimes b_i'' \in A \otimes B''$, then there exist $b_i \in B$ with $pb_i = b_i''$ for all i, for p is surjective. But

$$1 \otimes p : \sum a_i \otimes b_i \mapsto \sum a_i \otimes pb_i = \sum a_i \otimes b_i''. \quad \bullet$$

A similar statement holds for the functor $\quad \otimes_R B$. If B is a left R-module and

$$A' \xrightarrow{i} A \xrightarrow{p} A'' \to 0$$

is a short exact sequence of right R-modules, then the sequence

$$A' \otimes_R B \xrightarrow{i \otimes 1_B} A \otimes_R B \xrightarrow{p \otimes 1_B} A'' \otimes_R B \to 0$$

is exact.

Definition. A (covariant) functor $T: {}_R\mathbf{Mod} \to \mathbf{Ab}$ is called ***right exact*** if exactness of a sequence of left R-modules

$$B' \xrightarrow{i} B \xrightarrow{p} B'' \to 0$$

implies exactness of the sequence

$$T(B') \xrightarrow{T(i)} T(B) \xrightarrow{T(p)} T(B'') \to 0.$$

There is a similar definition for covariant functors $\mathbf{Mod}_R \to \mathbf{Ab}$.

In this terminology, the functors $A \otimes_R$ and $\otimes_R B$ are right exact functors.

The next example illustrates the absence of "$0 \to$" in Theorem 8.90.

Example 8.91.

Consider the exact sequence of abelian groups

$$0 \to \mathbb{Z} \xrightarrow{i} \mathbb{Q} \to \mathbb{Q}/\mathbb{Z} \to 0,$$

where i is the inclusion. By right exactness, there is an exact sequence

$$\mathbb{I}_2 \otimes \mathbb{Z} \xrightarrow{1 \otimes i} \mathbb{I}_2 \otimes \mathbb{Q} \to \mathbb{I}_2 \otimes (\mathbb{Q}/\mathbb{Z}) \to 0$$

(in this proof, we abbreviate $\otimes_{\mathbb{Z}}$ to $\otimes$). Now $\mathbb{I}_2 \otimes \mathbb{Z} \cong \mathbb{I}_2$, by Proposition 8.86. On the other hand, if $a \otimes q$ is a generator of $\mathbb{I}_2 \otimes \mathbb{Q}$, then

$$a \otimes q = a \otimes (2q/2) = 2a \otimes (q/2) = 0 \otimes (q/2) = 0.$$

Therefore, $\mathbb{I}_2 \otimes \mathbb{Q} = 0$, and so $1 \otimes i$ cannot be an injection. ◂

The next proposition helps compute tensor products.

Proposition 8.92. *For every abelian group B, we have $\mathbb{I}_n \otimes_{\mathbb{Z}} B \cong B/nB$.*

Proof. If A is a finite cyclic group of order n, there is an exact sequence

$$0 \to \mathbb{Z} \xrightarrow{\mu_n} \mathbb{Z} \xrightarrow{p} A \to 0,$$

where μ_n is multiplication by n. Tensoring by an abelian group B gives exactness of

$$\mathbb{Z} \otimes_{\mathbb{Z}} B \xrightarrow{\mu_n \otimes 1_B} \mathbb{Z} \otimes_{\mathbb{Z}} B \xrightarrow{p \otimes 1_B} A \otimes_{\mathbb{Z}} B \to 0.$$

Consider the diagram

$$\begin{array}{ccccccc} \mathbb{Z} \otimes_{\mathbb{Z}} B & \xrightarrow{\mu_n \otimes 1_B} & \mathbb{Z} \otimes_{\mathbb{Z}} B & \xrightarrow{p \otimes 1_B} & A \otimes_{\mathbb{Z}} B & \longrightarrow & 0 \\ \downarrow{\scriptstyle\theta} & & \downarrow{\scriptstyle\theta} & & & & \\ B & \xrightarrow[\mu_n]{} & B & \xrightarrow[\pi]{} & B/nB & \longrightarrow & 0, \end{array}$$

where $\theta\colon \mathbb{Z} \otimes_{\mathbb{Z}} B \to B$ is the isomorphism of Proposition 8.86, namely, $\theta\colon m \otimes b \mapsto mb$, where $m \in \mathbb{Z}$ and $b \in B$. This diagram commutes, for both composites take $m \otimes b \mapsto nmb$. The next, very general, proposition will apply to this diagram, yielding

$$A \otimes_{\mathbb{Z}} B \cong B/nB. \quad \bullet$$

Proposition 8.93. *Given a commutative diagram with exact rows in which the vertical maps f and g are isomorphisms,*

$$\begin{array}{ccccccc}
A' & \xrightarrow{i} & A & \xrightarrow{p} & A'' & \longrightarrow & 0 \\
{\scriptstyle f}\downarrow & & \downarrow{\scriptstyle g} & & \vdots\,{\scriptstyle h} & & \\
B' & \xrightarrow{j} & B & \xrightarrow{q} & B'' & \longrightarrow & 0,
\end{array}$$

there exists a unique isomorphism $h\colon A'' \to B''$ making the augmented diagram commute.

Proof. If $a'' \in A''$, then there is $a \in A$ with $p(a) = a''$ because p is surjective. Define $h(a'') = qg(a)$. Of course, we must show that h is well-defined; that is, if $u \in A$ satisfies $p(u) = a''$, then $qg(u) = qg(a)$. Since $p(a) = p(u)$, we have $p(a-u) = 0$, so that $a - u \in \ker p = \operatorname{im} i$, by exactness. Hence, $a - u = i(a')$, for some $a' \in A'$. Thus,

$$qg(a-u) = qgi(a') = qjf(a') = 0,$$

because $qj = 0$. Therefore, h is well-defined.

To see that the map h is an isomorphism, we construct its inverse. As in the first paragraph, there is a map h' making the following diagram commute:

$$\begin{array}{ccccccc}
B' & \xrightarrow{j} & B & \xrightarrow{q} & B'' & \longrightarrow & 0 \\
{\scriptstyle f^{-1}}\downarrow & & \downarrow{\scriptstyle g^{-1}} & & \vdots\,{\scriptstyle h'} & & \\
A' & \xrightarrow{i} & A & \xrightarrow{p} & A'' & \longrightarrow & 0
\end{array}$$

We claim that $h' = h^{-1}$. Now $h'q = pg^{-1}$. Hence,

$$h'hp = h'qg = pg^{-1}g = p;$$

since p is surjective, we have $h'h = 1_{A''}$. A similar calculation shows that the other composite hh' is also the identity. Therefore, h is an isomorphism. If $h'\colon A'' \to B''$ satisfies $h'p = qg$ and if $a'' \in A''$, choose $a \in A$ with $pa = a''$. Then $h'pa = h'a'' = qga = ha''$, and so h is unique. $\bullet$

The proof of the last proposition is an example of ***diagram chasing***. Such proofs appear long, but they are, in truth, quite routine. We select an element and, at each step, there is essentially only one thing to do with it. The proof of the dual proposition is another example of this sort of thing.

Proposition 8.94. *Given a commutative diagram with exact rows in which the vertical maps g and h are isomorphisms,*

$$\begin{array}{ccccccc} 0 & \longrightarrow & A' & \xrightarrow{i} & A & \xrightarrow{p} & A'' \\ & & \downarrow{\scriptstyle f} & & \downarrow{\scriptstyle g} & & \downarrow{\scriptstyle h} \\ 0 & \longrightarrow & B' & \xrightarrow{j} & B & \xrightarrow{q} & B'', \end{array}$$

there exists a unique isomorphism $f: A' \to B'$ making the augmented diagram commute.

Proof. A diagram chase. •

A tensor product of two nonzero modules can be zero. The following proposition generalizes the computation in Example 8.91.

Proposition 8.95. *If T is an abelian group with every element of finite order and if D is a divisible abelian group, then $T \otimes_{\mathbb{Z}} D = \{0\}$.*

Proof. It suffices to show that each generator $t \otimes d$, where $t \in T$ and $d \in D$, is 0 in $T \otimes_{\mathbb{Z}} D$. Since t has finite order, there is a nonzero integer n with $nt = 0$. As D is divisible, there exists $d' \in D$ with $d = nd'$. Hence,

$$t \otimes d = t \otimes nd' = nt \otimes d' = 0 \otimes d' = 0. \quad \bullet$$

We now understand why we cannot make a finite cyclic group G into a $\mathbb{Q}$-module, for $\mathbb{Q} \otimes_{\mathbb{Z}} G = \{0\}$.

Corollary 8.96. *If D is a nonzero divisible abelian group with every element of finite order* (*e.g., $D = \mathbb{Q}/\mathbb{Z}$*), *then there is no multiplication $D \times D \to D$ making D a ring.*

Proof. Assume, on the contrary, that there is a multiplication $\mu: D \times D \to D$ making D a ring. If 1 is the identity, we have $1 \neq 0$, lest D be the zero ring, which has only one element. Since multiplication in a ring is $\mathbb{Z}$-bilinear, there is a homomorphism $\widetilde{\mu}: D \otimes_{\mathbb{Z}} D \to D$ with $\widetilde{\mu}(d \otimes d') = \mu(d, d')$ for all $d, d' \in D$. In particular, if $d \neq 0$, then $\widetilde{\mu}(d \otimes 1) = \mu(d, 1) = d \neq 0$. But $D \otimes_{\mathbb{Z}} D = \{0\}$, by Proposition 8.95, so that $\widetilde{\mu}(d \otimes 1) = 0$. This contradiction shows that no multiplication μ on D exists. •

The next modules arise from tensor products in the same way that projective and injective modules arise from Hom. Investigation of the kernel of $A \otimes_R B' \to A \otimes_R B$ is done in homological algebra; it is intimately related with a functor called Tor.

Definition. If R is a ring, then a right R-module A is ***flat***[8] if, whenever

$$0 \to B' \xrightarrow{i} B \xrightarrow{p} B'' \to 0$$

[8] This term arose as the translation into algebra of a geometric property of varieties.

is an exact sequence of left R-modules, then

$$0 \to A \otimes_R B' \xrightarrow{1_A \otimes i} A \otimes_R B \xrightarrow{1_A \otimes p} A \otimes_R B'' \to 0$$

is an exact sequence of abelian groups. Flatness of a left R-module is defined similarly.

In other words, A is flat if and only if $A \otimes_R \ \ $ is an exact functor. Because the functor $A \otimes_R \ \ $ is right exact, we see that A is flat if and only if, whenever $i : B' \to B$ is an injection, then $1_A \otimes i : A \otimes_R B' \to A \otimes_R B$ is also an injection.

Lemma 8.97. *If every finitely generated submodule of a right R-module M is flat, then M is flat.*

Remark. Another proof of this lemma is given in Corollary 8.103. ◂

Proof. Let $i : A \to B$ be an injective R-map between left R-modules, and assume that $u = \sum_j x_j \otimes y_j \in \ker(1_M \otimes i)$, where $x_j \in M$ and $y_j \in A$. As $u \in M \otimes_R A$, we have

$$0 = (1_M \otimes i)u = \sum_{j=1}^{n} x_j \otimes i y_j.$$

Let F be the free abelian group with basis $M \times A$, and let S be the subgroup of F consisting of the relations of $F/S \cong M \otimes_R A$ (as in the construction of the tensor product in Proposition 8.74); thus, S is generated by all elements in F of the form

$$(m, a + a') - (m, a) - (m, a');$$
$$(m + m', a) - (m, a) - (m', a);$$
$$(mr, a) - (m, ra).$$

Let M' be the submodule of M generated by $x_1, \ldots, x_n$ together with the (finite number of) first "coordinates" in M exhibiting $\sum_k (x_j, i y_j)$ as a linear combination of relators just displayed. Of course, M' is a finitely generated submodule of M. The element $u' = \sum x_j \otimes y_j \in M' \otimes_R A$ (which is the version of u lying in this new tensor product $M' \otimes_R A$) lies in $\ker 1_{M'} \otimes i$, for we have taken care that all the relations making $(1_M \otimes i)(u) = 0$ are still present. But M' is a finitely generated submodule of M, so that it is flat, by hypothesis, and so $(1_{M'} \otimes i)(u) = 0$ implies $u' = 0$ in $M' \otimes_R A$. Finally, if $\ell : M' \to M$ is the inclusion, then $(\ell \otimes 1_A)(u') = u$, and so $u = 0$. Therefore, $1_M \otimes i$ is injective and M is flat. •

We will use this lemma to prove that an abelian group is a flat $\mathbb{Z}$-module if and only if it has no nonzero elements of finite order (see Corollary 9.6). Here are some examples of flat modules.

Lemma 8.98. *Let R be an arbitrary ring.*

(i) *The right R-module R is a flat R-module.*

(ii) *A direct sum of right R-modules $\sum_j M_j$ is flat if and only each M_j is flat.*

(iii) *Every projective right R-module F is flat.*

Proof. (i) Consider the commutative diagram

$$\begin{array}{ccc} A & \xrightarrow{\ i\ } & B \\ {\scriptstyle\sigma}\downarrow & & \downarrow{\scriptstyle\tau} \\ R \otimes_R A & \xrightarrow[1_R \otimes i]{} & R \otimes_R B \end{array}$$

where $i: A \to B$ is an injection, $\sigma: a \mapsto 1 \otimes a$, and $\tau: b \mapsto 1 \otimes b$. Now both σ and τ are isomorphisms, by Proposition 8.86, and so $1_R \otimes i = \tau i \sigma^{-1}$ is an injection. Therefore, R is a flat module over itself.

(ii) By Proposition 7.30, any family of R-maps $\{f_j: U_j \to V_j\}$ can be assembled into an R-map $\varphi: \sum_j U_j \to \sum_j V_j$, where $\varphi: (u_j) \mapsto (f_j(u_j))$, and it is easy to check that φ is an injection if and only if each f_j is an injection.

Let $i: A \to B$ be an injection. There is a commutative diagram

$$\begin{array}{ccc} (\sum_j M_j) \otimes_R A & \xrightarrow{1 \otimes i} & (\sum_j M_j) \otimes_R B \\ \downarrow & & \downarrow \\ \sum_j (M_j \otimes_R A) & \xrightarrow{\ \varphi\ } & \sum_j (M_j \otimes_R B), \end{array}$$

where $\varphi: (m_j \otimes a) \mapsto (m_j \otimes ia)$, where 1 is the identity map on $\sum_j M_j$, and where the downward maps are the isomorphisms of Proposition 8.87.

By our initial observation, $1 \otimes i$ is an injection if and only if each $1_{M_j} \otimes i$ is an injection; this says that $\sum_j M_j$ is flat if and only if each M_j is flat.

(iii) Combining the first two parts, we see that a free R-module, being a direct sum of copies of R, must be flat. Moreover, since a module is projective if and only if it is a direct summand of a free module, part (ii) shows that projective modules are always flat. •

We cannot improve this lemma without further assumptions, for there exist rings R for which every flat R-module is projective.

There is a remarkable relationship between Hom and $\otimes$. The key idea is that a function of two variables, say, $f: A \times B \to C$, can be viewed as a one-parameter family of functions of one variable: If we fix $a \in A$, then define $f_a: B \to C$ by $b \mapsto f(a, b)$. Recall Lemma 8.80: If R and S are rings and A_R and ${}_R B_S$ are modules, then $A \otimes_R B$ is a right S-module, where $(a \otimes b)s = a \otimes (bs)$. Furthermore, if C_S is a module, then it is easy to see that $\mathrm{Hom}_S(B, C)$ is a right R-module, where $(fr)(b) = f(rb)$; thus $\mathrm{Hom}_R(A, \mathrm{Hom}_S(B, C))$ makes sense, for it consists of R-maps between right R-modules. Finally, if $F \in \mathrm{Hom}_R(A, \mathrm{Hom}_S(B, C))$, we denote its value on $a \in A$ by F_a, so that $F_a: B \to C$, defined by $F_a: b \mapsto F(a)(b)$, is a one-parameter family of functions.

Theorem 8.99 (Adjoint Isomorphism). *Given modules A_R, ${}_RB_S$, and C_S, where R and S are rings, there is an isomorphism*

$$\tau_{A,B,C} \colon \operatorname{Hom}_S(A \otimes_R B, C) \to \operatorname{Hom}_R(A, \operatorname{Hom}_S(B, C)),$$

namely, for $f \colon A \otimes_R B \to C$ and $a \in A$ and $b \in B$,

$$\tau_{A,B,C} \colon f \mapsto f^*, where\ f_a^* \colon b \mapsto f(a \otimes b).$$

Indeed, fixing any two of A, B, C, the maps $\tau_{A,B,C}$ constitute natural equivalences

$$\operatorname{Hom}_S(\ \otimes_R B, C) \to \operatorname{Hom}_R(\ , \operatorname{Hom}_S(B, C)),$$

$$\operatorname{Hom}_S(A \otimes_R\ , C) \to \operatorname{Hom}_R(A, \operatorname{Hom}_S(\ , C)),$$

and

$$\operatorname{Hom}_S(A \otimes_R B,\ ) \to \operatorname{Hom}_R(A, \operatorname{Hom}_S(B,\ )).$$

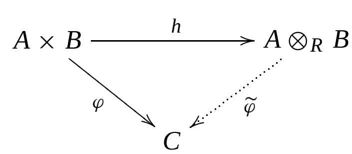
Proof. To prove that $\tau = \tau_{A,B,C}$ is a $\mathbb{Z}$-homomorphism, let $f, g \colon A \otimes_R B \to C$. The definition of $f + g$ gives, for all $a \in A$,

$$\begin{aligned} \tau(f+g)_a \colon b \mapsto (f+g)(a \otimes b) &= f(a \otimes b) + g(a \otimes b) \\ &= \tau(f)_a(b) + \tau(g)_a(b). \end{aligned}$$

Therefore, $\tau(f+g) = \tau(f) + \tau(g)$.

Next, τ is injective. If $\tau(f)_a = 0$ for all $a \in A$, then $0 = \tau(f)_a(b) = f(a \otimes b)$ for all $a \in A$ and $b \in B$. Therefore, $f = 0$ because it vanishes on every generator of $A \otimes_R B$.

We now show that τ is surjective. If $F \colon A \to \operatorname{Hom}_S(B, C)$ is an R-map, define $\varphi \colon A \times B \to C$ by $\varphi(a, b) = F_a(b)$. Now consider the diagram



It is straightforward to check that φ is R-biadditive, and so there exists a $\mathbb{Z}$-homomorphism $\widetilde{\varphi} \colon A \otimes_R B \to C$ with $\widetilde{\varphi}(a \otimes b) = \varphi(a, b) = F_a(b)$ for all $a \in A$ and $b \in B$. Therefore, $F = \tau(\widetilde{\varphi})$, so that τ is surjective.

We let the reader prove that the indicated maps are natural transformations; diagrams and the proof of their commutativity must be given. •

Given any two functors $F \colon \mathcal{C} \to \mathcal{D}$ and $G \colon \mathcal{D} \to \mathcal{C}$, we called the ordered pair (F, G) an ***adjoint pair*** if, for each pair of objects $C \in \mathcal{C}$ and $D \in \mathcal{D}$, there are bijections

$$\tau_{C,D} \colon \operatorname{Hom}_{\mathcal{D}}(FC, D) \to \operatorname{Hom}_{\mathcal{C}}(C, GD)$$

that are natural transformations in C and in D. It follows from Theorem 8.99 that $(\ \otimes_R B, \operatorname{Hom}(B,\ ))$ is an adjoint pair.

As promised earlier, here is another proof of Theorem 8.90, the right exactness of tensor product. Since $(\ \otimes_R B, \mathrm{Hom}(B\ ,\ ))$ is an adjoint pair of functors, the right functor $\otimes$ must preserve all direct limits, by Theorem 7.105. But cokernel is a direct limit, and a functor is right exact if it preserves cokernels. Here is this proof in more concrete terms.

Proposition 8.100. *Let A be a right R-module, and let*

$$B' \xrightarrow{i} B \xrightarrow{p} B'' \to 0$$

be an exact sequence of left R-modules. Then

$$A \otimes_R B' \xrightarrow{1_A \otimes i} A \otimes_R B \xrightarrow{1_A \otimes p} A \otimes_R B'' \to 0$$

is an exact sequence of abelian groups.

Proof. Regard a left R-module B as a $(R, \mathbb{Z})$-bimodule, and note, for any abelian group C, that $\mathrm{Hom}_{\mathbb{Z}}(B, C)$ is a right R-module, by Exercise 8.45 on page 603. In light of Proposition 7.48, it suffices to prove that the top row of the following diagram is exact for every C:

$$\begin{array}{ccccccc}
0 \to & \mathrm{Hom}_{\mathbb{Z}}(A \otimes_R B'', C) & \to & \mathrm{Hom}_{\mathbb{Z}}(A \otimes_R B, C) & \to & \mathrm{Hom}_{\mathbb{Z}}(A \otimes_R B', C) \\
& \downarrow \tau''_{A,C} & & \downarrow \tau_{A,C} & & \downarrow \tau'_{A,C} \\
0 \longrightarrow & \mathrm{Hom}_R(A, H'') & \longrightarrow & \mathrm{Hom}_R(A, H) & \longrightarrow & \mathrm{Hom}_R(A, H'),
\end{array}$$

where $H'' = \mathrm{Hom}_{\mathbb{Z}}(B'', C)$, $H = \mathrm{Hom}_{\mathbb{Z}}(B, C)$, and $H' = \mathrm{Hom}_{\mathbb{Z}}(B', C)$. By the adjoint isomorphism, the vertical maps are isomorphisms and the diagram commutes. The bottom row is exact, for it arises from the given exact sequence $B' \to B \to B'' \to 0$ by first applying the left exact (contravariant) functor $\mathrm{Hom}_{\mathbb{Z}}(\ , C)$, and then applying the left exact (covariant) functor $\mathrm{Hom}_R(A,\ )$. Exactness of the top row now follows from Exercise 8.51 on page 604. •

In Theorem 7.92, we proved that $\mathrm{Hom}(A,\ )$ preserves inverse limits; we now prove that $A\otimes$ preserves direct limits. This, too, follows from Theorem 7.105. However, we give another proof based on the construction of direct limits.

Theorem 8.101. *If A is a right R-module and $\{B_i, \varphi^i_j\}$ is a direct system of left R-modules* (*over any not necessarily directed index set I*), *then*

$$A \otimes_R \varinjlim B_i \cong \varinjlim (A \otimes_R B_i).$$

Proof. Note that Exercise 7.66 on page 517 shows that $\{A \otimes_R B_i, 1 \otimes \varphi^i_j\}$ is a direct system, so that $\varinjlim (A \otimes_R B_i)$ makes sense.

We begin by constructing $\varinjlim B_i$ as the cokernel of a certain map between sums. For each pair $i, j \in I$ with $i \preceq j$ in the partially ordered index set I, define B_{ij} to be a module isomorphic to B_i by a map $b_i \mapsto b_{ij}$, where $b_i \in B_i$, and define $\sigma\colon \sum_{ij} B_{ij} \to \sum_i B_i$ by

$$\sigma\colon b_i \mapsto \lambda_j \varphi^i_j b_i - \lambda_i b_i,$$

where λ_i is the injection of B_i into the sum. Note that $\operatorname{im}\sigma = S$, the submodule arising in the construction of $\varinjlim B_i$ in Proposition 7.94. Thus, $\operatorname{coker}\sigma = (\sum B_i)/S \cong \varinjlim B_i$, and there is an exact sequence

$$\sum B_{ij} \xrightarrow{\sigma} \sum B_i \to \varinjlim B_i \to 0.$$

Right exactness of $A\otimes_R \;\;$ gives exactness of

$$A \otimes_R \Big(\sum B_{ij}\Big) \xrightarrow{1\otimes\sigma} A \otimes_R \Big(\sum B_i\Big) \to A \otimes_R (\varinjlim B_i) \to 0.$$

By Theorem 8.87, the map $\tau\colon A \otimes_R \Big(\sum_i B_i\Big) \to \sum_i (A \otimes_R B_i)$, given by

$$\tau\colon a \otimes (b_i) \mapsto (a \otimes b_i),$$

is an isomorphism, and so there is a commutative diagram

$$\begin{array}{ccccccc} A \otimes \sum B_{ij} & \xrightarrow{1\otimes\sigma} & A \otimes \sum B_i & \longrightarrow & A \otimes \varinjlim B_i & \longrightarrow & 0 \\ \downarrow{\scriptstyle\tau} & & \downarrow{\scriptstyle\tau'} & & \vdots & & \\ \sum(A \otimes B_{ij}) & \xrightarrow{\widetilde{\sigma}} & \sum(A \otimes B_i) & \longrightarrow & \varinjlim(A \otimes B_i) & \longrightarrow & 0, \end{array}$$

where τ' is another instance of the isomorphism of Theorem 8.87, and

$$\widetilde{\sigma}\colon a \otimes b_{ij} \mapsto (1 \otimes \lambda_j)(a \otimes \varphi^i_j b_i) - (1 \otimes \lambda_i)(a' \otimes b_i).$$

By Proposition 8.93, there is an isomorphism $A \otimes_R \varinjlim B_i \to \operatorname{coker}\widetilde{\sigma} \cong \varinjlim(A \otimes_R B_i)$, the direct limit of the direct system $\{A \otimes_R B_i, 1 \otimes \varphi^i_j\}$. •

The reader has probably observed that we have actually proved a stronger result: any right exact functor that preserves sums must preserve all direct limits. The dual result also holds, and it has a similar proof; every left exact functor that preserves products must preserve all inverse limits. In fact, if (F, G) is an adjoint pair of functors (defined on module categories), then F preserves direct limits and G preserves inverse limits.

Corollary 8.102. *If $\{F_i, \varphi^i_j\}$ is a direct system of flat right R-modules over a directed index set I, then $\varinjlim F_i$ is also flat.*

Proof. Let $0 \to A \xrightarrow{k} B$ be an exact sequence of left R-modules. Since each F_i is flat, the sequence

$$0 \to F_i \otimes_R A \xrightarrow{1_i\otimes k} F_i \otimes_R B$$

is exact for every i, where 1_i abbreviates 1_{F_i}. Consider the commutative diagram

$$\begin{array}{ccccc} 0 & \longrightarrow & \varinjlim(F_i \otimes A) & \xrightarrow{\vec{k}} & \varinjlim(F_i \otimes B) \\ & & \downarrow{\scriptstyle\varphi} & & \downarrow{\scriptstyle\psi} \\ 0 & \longrightarrow & (\varinjlim F_i) \otimes A & \xrightarrow{1\otimes k} & (\varinjlim F_i) \otimes B, \end{array}$$

where the vertical maps φ and ψ are the isomorphisms of Theorem 8.101, the map $\vec{k}$ is induced from the transformation of direct systems $\{1_i \otimes k\}$, and 1 is the identity map on $\varinjlim F_i$. Since each F_i is flat, the maps $1_i \otimes k$ are injections; since the index set I is directed, the top row is exact, by Proposition 7.100. Therefore, $1 \otimes k \colon (\varinjlim F_i) \otimes A \to (\varinjlim F_i) \otimes B$ is an injection, for it is the composite of injections $\psi \vec{k} \varphi^{-1}$. Therefore, $\varinjlim F_i$ is flat. •

Corollary 8.103.

(i) *If R is a domain with $Q = \mathrm{Frac}(R)$, then Q is a flat R-module.*

(ii) *If every finitely generated submodule of a right R-module M is flat, then M is flat.*

Proof. (i) In Example 7.97(v), we saw that Q is a direct limit, over a directed index set, of cyclic submodules, each of which is isomorphic to R. Since R is projective, hence flat, the result follows from Corollary 8.102.

(ii) In Example 7.99(iii), we saw that M is a direct limit, over a directed index set, of its finitely generated submodules. Since every finitely generated submodule is flat, by hypothesis, the result follows from Corollary 8.102. We have given another proof of Lemma 8.97. •

Corollary 7.75 can be extended from abelian groups to modules over any ring.

Theorem 8.104. *For every ring R, every left R-module M can be imbedded as a submodule of an injective left R-module.*

Proof. Regarding R as a bimodule ${}_{\mathbb{Z}}R_R$ and an abelian group D as a left $\mathbb{Z}$-module, we use Exercise 8.45 on page 603 to see that $\mathrm{Hom}_{\mathbb{Z}}(R, D)$ is a left R-module; the scalar multiplication $R \times \mathrm{Hom}_{\mathbb{Z}}(R, D) \to \mathrm{Hom}_{\mathbb{Z}}(R, D)$ is given by $(a, \varphi) \mapsto a\varphi$, where $a\varphi \colon r \mapsto \varphi(ra)$.

If now D is a divisible abelian group, we claim that $H = \mathrm{Hom}_{\mathbb{Z}}(R, D)$ is an injective R-module; that is, we show that $\mathrm{Hom}_R(\quad, H)$ is an exact functor. Since Hom is left exact, it suffices to show that if $i \colon A' \to A$ is an injection, then the induced map $i^* \colon \mathrm{Hom}_R(A, H) \to \mathrm{Hom}_R(A', H)$ is a surjection. Consider the following diagram.

$$
\begin{array}{ccc}
\mathrm{Hom}_R(A, \mathrm{Hom}_{\mathbb{Z}}(R, D)) & \xrightarrow{i^*} & \mathrm{Hom}_R(A', \mathrm{Hom}_{\mathbb{Z}}(R, D)) \\
\downarrow & & \downarrow \\
\mathrm{Hom}_{\mathbb{Z}}(A \otimes_R R, D) & \longrightarrow & \mathrm{Hom}_{\mathbb{Z}}(A' \otimes_R R, D) \\
\downarrow & & \downarrow \\
\mathrm{Hom}_{\mathbb{Z}}(A, D) & \longrightarrow & \mathrm{Hom}_{\mathbb{Z}}(A', D)
\end{array}
$$

The adjoint isomorphism gives commutativity of the top square. The bottom square arises from applying the contravariant functor $\mathrm{Hom}_{\mathbb{Z}}(\quad, D)$ to the following diagram, which

commutes because the isomorphism $A \to A \otimes_R R$, given by $a \mapsto a \otimes 1$, is natural.

$$\begin{array}{ccc} A \otimes_R R & \longleftarrow & A' \otimes_R R \\ \uparrow & & \uparrow \\ A & \longleftarrow & A' \end{array}$$

Since D is divisible, Corollary 7.73 says that D is an injective $\mathbb{Z}$-module. Therefore, $\text{Hom}_{\mathbb{Z}}(\quad, D)$ is an exact functor and the bottom row in the large diagram is surjective. Since all the vertical maps in the large diagram are isomorphisms, commutativity now gives i^* surjective. We conclude that $\text{Hom}_{\mathbb{Z}}(R, D)$ is an injective left R-module.

Finally, regard M as an abelian group. By Corollary 7.75, there is a divisible abelian group D and an injective $\mathbb{Z}$-homomorphism $j : M \to D$. It is now easy to see that there is an injective R-map $M \to \text{Hom}_{\mathbb{Z}}(R, D)$, namely, $m \mapsto f_m$, where $f_m(r) = j(rm) \in D$; this completes the proof. •

This last theorem can be improved, for there is a smallest injective module containing any given module, called its ***injective envelope*** (see Rotman, *An Introduction to Homological Algebra*, page 73).

We have already seen, in Proposition 7.69, that if R is a noetherian ring, then every direct sum of injective modules is injective; we now prove the converse.

Theorem 8.105 (*Bass*). *If R is a ring for which every direct sum of injective left R-modules is injective, then R is left noetherian.*

Proof. We show that if R is not left noetherian, then there is a left ideal I and an R-map to a sum of injectives that cannot be extended to R. Since R is not left noetherian, there is a strictly ascending chain of left ideals $I_1 \subsetneq I_2 \subsetneq \cdots$; let $I = \bigcup I_n$. We note that $I/I_n \neq \{0\}$ for all n. By Theorem 8.104, we may imbed I/I_n in an injective left R-module E_n; we claim that $E = \sum_n E_n$ is not injective.

Let $\pi_n : I \to I/I_n$ be the natural map. For each $a \in I$, note that $\pi_n(a) = 0$ for large n (because $a \in I_n$ for some n), and so the R-map $f : I \to \prod(I/I_n)$, defined by

$$f : a \mapsto (\pi_n(a)),$$

does have its image in $\sum_n(I/I_n)$; that is, for each $a \in I$, almost all the coordinates of $f(a)$ are 0. Composing with the inclusion $\sum(I/I_n) \to \sum E_n = E$, we may regard f as a map $I \to E$. If there is an R-map $g : R \to E$ extending f, then $g(1)$ is defined; say, $g(1) = (x_n)$. Choose an index m and choose $a \in I$ with $a \notin I_m$; since $a \notin I_m$, we have $\pi_m(a) \neq 0$, and so $g(a) = f(a)$ has nonzero mth coordinate $\pi_m(a)$. But $g(a) = ag(1) = a(x_n) = (ax_n)$, so that $\pi_m(a) = ax_m$. It follows that $x_m \neq 0$ for all m, and this contradicts $g(1)$ lying in the direct sum $E = \sum E_n$. •

We are now going to give a connection between flat modules and projective modules.

Definition. If B is a right R-module, define its ***character module*** B^* as the left R-module

$$B^* = \operatorname{Hom}_{\mathbb{Z}}(B, \mathbb{Q}/\mathbb{Z}).$$

Recall that B^* is a left R-module if one defines rf, for $r \in R$ and $f : B \to \mathbb{Q}/\mathbb{Z}$, by

$$rf : b \mapsto f(br).$$

The next lemma improves Proposition 7.48: If $i : A' \to A$ and $p : A \to A''$ are maps and, for every module B,

$$0 \to \operatorname{Hom}(A'', B) \xrightarrow{p^*} \operatorname{Hom}(A, B) \xrightarrow{i^*} \operatorname{Hom}(A', B)$$

is an exact sequence, then so is

$$A' \xrightarrow{i} A \xrightarrow{p} A'' \to 0.$$

Lemma 8.106. *A sequence of right R-modules*

$$0 \to A \xrightarrow{\alpha} B \xrightarrow{\beta} C \to 0$$

is exact if and only if the sequence of character modules

$$0 \to C^* \xrightarrow{\beta^*} B^* \xrightarrow{\alpha^*} A^* \to 0$$

is exact.

Proof. If the original sequence is exact, then so is the sequence of character modules, for the contravariant functor $\operatorname{Hom}_{\mathbb{Z}}(\ , \mathbb{Q}/\mathbb{Z})$ is exact, because $\mathbb{Q}/\mathbb{Z}$ is an injective $\mathbb{Z}$-module, by Corollary 7.73.

For the converse, it suffices to prove that $\ker\alpha = \operatorname{im}\beta$ without assuming either α^* surjective or β^* is injective.

$\operatorname{im}\alpha \subseteq \ker\beta$.

If $x \in A$ and $\alpha x \notin \ker\beta$, then $\beta\alpha(x) \neq 0$. By Exercise 7.57(i) on page 488, there is a map $f : C \to \mathbb{Q}/\mathbb{Z}$ with $f\beta\alpha(x) \neq 0$. Thus, $f \in C^*$ and $f\beta\alpha \neq 0$, which contradicts the hypothesis that $\alpha^*\beta^* = 0$.

$\ker\beta \subseteq \operatorname{im}\alpha$.

If $y \in \ker\beta$ and $y \notin \operatorname{im}\alpha$, then $y + \operatorname{im}\alpha$ is a nonzero element of $B/\operatorname{im}\alpha$. Thus, there is a map $g : B/\operatorname{im}\alpha \to \mathbb{Q}/\mathbb{Z}$ with $g(y + \operatorname{im}\alpha) \neq 0$, by Exercise 7.57(i) on page 488. If $\nu : B \to B/\operatorname{im}\alpha$ is the natural map, define $g' = g\nu \in B^*$; note that $g'(y) \neq 0$, for $g'(y) = g\nu(y) = g(y + \operatorname{im}\alpha)$. Now $g'(\operatorname{im}\alpha) = \{0\}$, so that $0 = g'\alpha = \alpha^*(g')$ and $g' \in \ker\alpha^* = \operatorname{im}\beta^*$. Thus, $g' = \beta^*(h)$ for some $h \in C^*$; that is, $g' = h\beta$. Hence, $g'(y) = h\beta(y)$, which is a contradiction, for $g'(y) \neq 0$, while $h\beta(y) = 0$, because $y \in \ker\beta$. •

Proposition 8.107. *A right R-module B is flat if and only if its character module B^* is an injective left R-module.*

Proof. As in the proof of Theorem 8.104 with B playing the role of R (so that flatness implies that the map $A' \otimes_R B \to A \otimes_R B$ is injective), the left R-module $B^* = \text{Hom}_{\mathbb{Z}}(B, \mathbb{Q}/\mathbb{Z})$ is injective.

Conversely, assume that B^* is an injective left R-module and $A' \to A$ is an injection between left R-modules A' and A. Since $\text{Hom}_R(A, B^*) = \text{Hom}_R(A, \text{Hom}_{\mathbb{Z}}(B, \mathbb{Q}/\mathbb{Z}))$, the adjoint isomorphism, Theorem 8.99, gives a commutative diagram in which the vertical maps are isomorphisms.

$$\begin{array}{ccccc}
\text{Hom}_R(A, B^*) & \longrightarrow & \text{Hom}_R(A', B^*) & \longrightarrow & 0 \\
\downarrow & & \downarrow & & \\
\text{Hom}_{\mathbb{Z}}(B \otimes_R A, \mathbb{Q}/\mathbb{Z}) & \longrightarrow & \text{Hom}_{\mathbb{Z}}(B \otimes A', \mathbb{Q}/\mathbb{Z}) & \longrightarrow & 0 \\
\| & & \| & & \\
(B \otimes_R A)^* & \longrightarrow & (B \otimes_R A')^* & \longrightarrow & 0
\end{array}$$

Exactness of the top row now gives exactness of the bottom row. By Lemma 8.106, the sequence $0 \to B \otimes_R A' \to B \otimes_R A$ is exact, and this gives B flat. •

Corollary 8.108. *A right R-module B is flat if and only if, for every finitely generated left ideal I, the sequence $0 \to B \otimes_R I \to B \otimes_R R$ is exact.*

Proof. If B is flat, then the sequence $0 \to B \otimes_R I \to B \otimes_R R$ is exact for every left R-module I; in particular, this sequence is exact when I is a finitely generated left ideal. Conversely, the hypothesis of exactness of $0 \to B \otimes_R I \to B \otimes_R R$ for every finitely generated left ideal I allows us to prove exactness of this sequence for every left ideal, using Proposition 7.100 and the fact that tensor product commutes with direct limits. There is an exact sequence $(B \otimes_R R)^* \to (B \otimes_R I)^* \to 0$ that, by the adjoint isomorphism, gives exactness of $\text{Hom}_R(R, B^*) \to \text{Hom}_R(I, B^*) \to 0$. This says that every map from an ideal I to B^* extends to a map $R \to B^*$; thus, B^* satisfies the Baer criterion, Theorem 7.68, and so B^* is injective. By Proposition 8.107, B is flat. •

Lemma 8.109. *Given modules $({}_R X, {}_R Y_S, Z_S)$, where R and S are rings, there is a natural transformation in X, Y, and Z*

$$\tau_{X,Y,Z} \colon \text{Hom}_S(Y, Z) \otimes_R X \to \text{Hom}_S(\text{Hom}_R(X, Y), Z).$$

Moreover, $\tau_{X,Y,Z}$ an isomorphism whenever X is a finitely generated free left R-module.

Proof. Note that both $\text{Hom}_S(Y, Z)$ and $\text{Hom}_R(X, Y)$ make sense, for Y is a bimodule. If $f \in \text{Hom}_S(Y, Z)$ and $x \in X$, define $\tau_{X,Y,Z}(f \otimes x)$ to be the S-map $\text{Hom}_R(X, Y) \to Z$ given by

$$\tau_{X,Y,Z}(f \otimes x) \colon g \mapsto f(g(x)).$$

It is straightforward to check that $\tau_{X,Y,Z}$ is a homomorphism natural in X, that $\tau_{R,Y,Z}$ is an isomorphism, and, more generally, that $\tau_{X,Y,Z}$ is an isomorphism when X is a finitely generated free left R-module. •

Theorem 8.110. *A finitely presented left R-module B is flat if and only if it is projective.*

Proof. All projective modules are flat, by Lemma 8.98, and so only the converse is significant. Since B is finitely presented, there is an exact sequence

$$F' \to F \to B \to 0,$$

where both F' and F are finitely generated free left R-modules. We begin by showing, for every left R-module Y [which is necessarily an $(R - \mathbb{Z})$-bimodule], that the map $\tau_B = \tau_{B,Y,\mathbb{Q}/\mathbb{Z}} \colon Y^* \otimes_R B \to \mathrm{Hom}_R(B, Y)^*$ of Lemma 8.109 is an isomorphism.

Consider the following diagram.

$$\begin{array}{ccccccc}
Y^* \otimes_R F' & \longrightarrow & Y^* \otimes_R F & \longrightarrow & Y^* \otimes_R B & \longrightarrow & 0 \\
\downarrow{\scriptstyle \tau_{F'}} & & \downarrow{\scriptstyle \tau_F} & & \downarrow{\scriptstyle \tau_B} & & \\
\mathrm{Hom}_R(F', Y)^* & \longrightarrow & \mathrm{Hom}_R(F, Y)^* & \longrightarrow & \mathrm{Hom}_R(B, Y)^* & \longrightarrow & 0
\end{array}$$

By Lemma 8.109, this diagram commutes [for $Y^* \otimes_R F = \mathrm{Hom}_{\mathbb{Z}}(Y, \mathbb{Q}/\mathbb{Z}) \otimes_R F$] and the first two vertical maps are isomorphisms. The top row is exact, because $Y^* \otimes_R \;$ is right exact. The bottom row is also exact, because $\mathrm{Hom}_R(\ , Y)^*$ is the composite of the contravariant functors $\mathrm{Hom}_R(\ , Y)$, which is left exact, and $^* = \mathrm{Hom}_{\mathbb{Z}}(\ , \mathbb{Q}/\mathbb{Z})$, which is exact. Proposition 8.93 now shows that the third vertical arrow, $\tau_B \colon Y^* \otimes_R B \to \mathrm{Hom}_R(B, Y)^*$, is an isomorphism.

To prove that B is projective, it suffices to prove that $\mathrm{Hom}(B, \ )$ preserves surjections: If $A \to A'' \to 0$ is exact, then $\mathrm{Hom}(B, A) \to \mathrm{Hom}(B, A'') \to 0$ is exact. By Lemma 8.106, it suffices to show that $0 \to \mathrm{Hom}(B, A'')^* \to \mathrm{Hom}(B, A)^*$ is exact. Consider the diagram

$$\begin{array}{ccccc}
0 & \longrightarrow & A''^* \otimes_R B & \longrightarrow & A^* \otimes_R B \\
 & & \downarrow{\scriptstyle \tau} & & \downarrow{\scriptstyle \tau} \\
0 & \longrightarrow & \mathrm{Hom}(B, A'')^* & \longrightarrow & \mathrm{Hom}(B, A)^*
\end{array}$$

Naturality of τ gives commutativity of the diagram, while the vertical maps τ are isomorphisms, because B is finitely presented. Since $A \to A'' \to 0$ is exact, $0 \to A''^* \to A^*$ is exact, and so the top row is exact, because B is flat. It follows that the bottom row is also exact; that is, $0 \to \mathrm{Hom}(B, A'')^* \to \mathrm{Hom}(B, A'')^*$ is exact, which is what we were to show. Therefore, B is projective. •

Corollary 8.111. *If R is right noetherian, then a finitely generated right R-module B is flat if and only if it is projective.*

Proof. This follows from the theorem once we recall Proposition 7.59 (which holds for noncommutative rings): Every finitely generated module over a noetherian ring is finitely presented. •

Here is a nice application of tensor products that helps to place the Wedderburn–Artin theorems in perspective.

Definition. A module P is ***small*** if the covariant Hom functor $\mathrm{Hom}(P, \quad)$ preserves (possibly infinite) direct sums.

For example, Proposition 8.85 shows that every ring R is a small R-module.

To say that P is small means more than that there is some isomorphism

$$\mathrm{Hom}\big(P, \sum_{i \in I} B_i\big) \cong \sum_{i \in I} \mathrm{Hom}(P, B_i);$$

it also means that $\mathrm{Hom}(P, \quad)$ preserves the coproduct diagram; if $\lambda_i \colon B_i \to B$ are the injections, where $B = \sum_{i \in I} B_i$, then the induced maps $(\lambda_i)_* \colon \mathrm{Hom}(P, B_i) \to \mathrm{Hom}(P, B)$ are the injections of $\sum_{i \in I} \mathrm{Hom}(P, B_i)$.

Example 8.112.

(i) Any finite direct sum of small modules is small, and any direct summand of a small module is small.

(ii) Since a ring R is a small R-module, it follows from (i) that every finitely generated free R-module is small and that every finitely generated projective R-module is small. ◂

Definition. A right R-module P is a ***generator of*** $\mathbf{Mod}_R$ if every right R-module M is a quotient of some direct sum of copies of P.

It is clear that R is a generator of $\mathbf{Mod}_R$, as is any free right R-module. However, a projective right R-module may not be a generator. For example, if $R = \mathbb{I}_6$, then $R = P \oplus Q$, where $P = \{[0], [2], [4]\} \cong \mathbb{I}_3$, and the projective module P is not a generator (for $Q \cong \mathbb{I}_2$ is not a quotient of a direct sum of copies of P).

Theorem 8.113. *Let R be a ring and let P be a small projective generator of* $\mathbf{Mod}_R$. *If $S = \mathrm{End}_R(P)$, then there is an equivalence of categories*

$$\mathbf{Mod}_R \cong \mathbf{Mod}_S.$$

Proof. Notice that P is a left S-module, for if $x \in P$ and $f, g \in S = \mathrm{End}_R(P)$, then $(g \circ f)x = g(fx)$. In fact, P is a (S, R)-bimodule, for associativity $f(xr) = (fx)r$, where $r \in R$, is just the statement that f is an R-map. It now follows from Corollary 8.81

that the functor $F\colon \mathbf{Mod}_S \to \mathbf{Ab}$, defined by $F = \quad \otimes_S P$, actually takes values in $\mathbf{Mod}_R$. Exercise 8.45(ii) on page 603 shows that the functor $G\colon\ \mathrm{Hom}_R(P,\quad)\colon \mathbf{Mod}_R \to \mathbf{Ab}$ actually takes values in $\mathbf{Mod}_S$. As (F, G) is an adjoint pair, Exercise 7.75 on page 518 gives natural transformations $FG \to 1_R$ and $1_S \to GF$, where 1_R and 1_S denote identity functors on the categories $\mathbf{Mod}_R$ and $\mathbf{Mod}_S$, respectively. It suffices to prove that each of these natural transformations is a natural equivalence.

Since P is a projective right R-module, the functor $G = \mathrm{Hom}_R(P,\quad)$ is exact; since P is small, G preserves direct sums. Now $F = \quad \otimes_S P$, as any tensor product functor, is right exact and preserve sums. Therefore, both composites GF and FG preserve direct sums and are right exact.

Note that

$$FG(P) = F(\mathrm{Hom}_R(P, P)) = F(S) = S \otimes_S P \cong P.$$

Since P is a generator of $\mathbf{Mod}_R$, every right R-module M is a quotient of some direct sum of copies of P: There is an exact sequence $K \to \sum P \xrightarrow{f} M \to 0$, where $K = \ker f$. There is also some direct sum of copies of P mapping onto K, and so there is an exact sequence

$$\sum P \to \sum P \to M \to 0.$$

Hence, there is a commutative diagram (by naturality of the upward maps) with exact rows

$$\begin{array}{ccccccc}
\sum P & \longrightarrow & \sum P & \longrightarrow & M & \longrightarrow & 0 \\
\uparrow & & \uparrow & & \uparrow & & \\
\sum FG(P) & \longrightarrow & \sum FG(P) & \longrightarrow & FG(M) & \longrightarrow & 0
\end{array}$$

We know that the first two vertical maps are isomorphisms, and so a diagram chase (see the five lemma, Exercise 8.52 on page 604) gives the other vertical map an isomorphism; that is, $FG(M) \cong M$, and so $1_R \cong FG$.

For the other composite, note that

$$GF(S) = G(S \otimes_S P) \cong G(P) = \mathrm{Hom}_R(P, P) = S.$$

If N is any left S-module, there is an exact sequence of the form

$$\sum S \to \sum S \to N \to 0,$$

because every module is a quotient of a free module. The argument now concludes as that just done. •

Corollary 8.114. *If R is a ring and $n \geq 1$, there is an equivalence of categories*

$$\mathbf{Mod}_R \cong \mathbf{Mod}_{\mathrm{Mat}_n(R)}.$$

Remark. There is an equivalence of categories ${}_{R^{\mathrm{op}}}\mathbf{Mod} \to \mathbf{Mod}_R$, by Exercise 8.22 on page 533. In particular, if R is commutative, then

$${}_R\mathbf{Mod} \cong \mathbf{Mod}_{\mathrm{Mat}_n(R)}. \quad ◀$$

Proof. For any integer $n \geq 1$, the free module $P = \sum_{i=1}^{n} R_i$, where $R_i \cong R$, is a small projective generator of $\mathbf{Mod}_R$, and $S = \mathrm{End}_R(P) \cong \mathrm{Mat}_n(R)$. •

We can now understand Proposition 8.49: $\mathrm{Mat}_n(\Delta)$ is semisimple when Δ is a division ring. By Proposition 8.48, a ring R is semisimple if and only if every R-module is projective; that is, every object in $\mathbf{Mod}_R$ is projective. Now every Δ-module is projective (even free), so that equivalence of the categories shows that every object in $\mathbf{Mod}_{\mathrm{Mat}_n(\Delta)}$ is also projective. Therefore, $\mathrm{Mat}_n(\Delta)$ is also semisimple.

There is a circle of ideas, usually called ***Morita theory*** (after K. Morita). The first question it asks is when an abstract category $\mathcal{C}$ is equivalent to $\mathbf{Mod}_R$ for some ring R. The answer is very nice: A category $\mathcal{C}$ is isomorphic to a module category if and only if it is an ***abelian category*** (this just means that the usual finitary constructions in the second section of Chapter 7 exist; see Mac Lane, *Categories for the Working Mathematician*, pages 187–206), it is closed under infinite coproducts, and it contains a small projective object P that is a generator. Given this hypothesis, then $\mathcal{C} \cong \mathbf{Mod}_S$, where $S = \mathrm{End}(P)$ (the proof is essentially that given for Theorem 8.113).

Two rings R and S are called ***Morita equivalent*** if $\mathbf{Mod}_R \cong \mathbf{Mod}_S$. For example, it follows from Theorem 8.113 that every commutative ring R is Morita equivalent to the ring $\mathrm{Mat}_n(R)$, where $n \geq 1$. Moreover, if R and S are Morita equivalent, then $Z(R) \cong Z(S)$; that is, they have isomorphic centers (the proof actually identifies all the possible isomorphisms between the categories). In particular, two commutative rings are Morita equivalent if and only if they are isomorphic. See Jacobson, *Basic Algebra* II, pages 177–184, Lam, *Lectures on Modules and Rings*, Chapters 18 and 19, and Reiner, *Maximal Orders*, Chapter 4.

In the next chapter, we will see that the tensor product $R \otimes_k S$ of two k-algebras R and S is also a k-algebra. Indeed, when R and S are commutative, then $R \otimes_k S$ is their coproduct in the category of commutative k-algebras.

EXERCISES

8.45 This exercise is an analog of Corollary 8.81.

(i) Given a bimodule ${}_RA_S$, prove that $\mathrm{Hom}_R(A, \ ) \colon {}_R\mathbf{Mod} \to {}_S\mathbf{Mod}$ is a functor, where $\mathrm{Hom}_R(A, B)$ is the left S-module defined by $sf \colon a \mapsto f(as)$.

(ii) Given a bimodule ${}_RA_S$, prove that $\mathrm{Hom}_S(A, \ ) \colon \mathbf{Mod}_S \to \mathbf{Mod}_R$ is a functor, where $\mathrm{Hom}_S(A, B)$ is the right R-module defined by $fr \colon a \mapsto f(ra)$.

(iii) Given a bimodule ${}_SB_R$, prove that $\mathrm{Hom}_R(\ , B) \colon \mathbf{Mod}_R \to {}_S\mathbf{Mod}$ is a functor, where $\mathrm{Hom}_R(A, B)$ is the left S-module defined by $sf \colon a \mapsto s[f(a)]$.

(iv) Given a bimodule ${}_SB_R$, prove that $\mathrm{Hom}_S(A, \ ) \colon {}_S\mathbf{Mod} \to \mathbf{Mod}_R$ is a functor, where $\mathrm{Hom}_S(A, B)$ is the right R-module defined by $fr \colon a \mapsto f(a)r$.

Remark. Let $f: A \to B$ be an R-map. Suppose we write $f(a)$ when A is a right R-module and $(a)f$ when A is a left R-module (that is, write the function symbol f on the side opposite the scalar action). With this notation, each of the four parts of this exercise is an associative law. For example, in part (i) with both A and B left R-modules, writing sf for $s \in S$, we have $a(sf) = (as)f$. Similarly, in part (ii), we define fr, for $r \in R$ so that $(fr)a = f(ra)$. ◂

8.46 Let V and W be finite-dimensional vector spaces over a field F, say, and let $v_1, \ldots, v_m$ and $w_1, \ldots, w_n$ be bases of V and W, respectively. Let $S: V \to V$ be a linear transformation having matrix $A = [a_{ij}]$, and let $T: W \to W$ be a linear transformation having matrix $B = [b_{k\ell}]$. Show that the matrix of $S \otimes T: V \otimes_k W \to V \otimes_k W$, with respect to a suitable listing of the vectors $v_i \otimes w_j$, is the $nm \times nm$ matrix K, which we write in block form:

$$A \otimes B = \begin{bmatrix} a_{11}B & a_{12}B & \cdots & a_{1m}B \\ a_{21}B & a_{22}B & \cdots & a_{2m}B \\ \vdots & \vdots & \vdots & \vdots \\ a_{m1}B & a_{m2}B & \cdots & a_{mm}B \end{bmatrix}.$$

The matrix $A \otimes B$ is called the ***Kronecker product*** of the matrices A and B.

8.47 Let R be a domain with $Q = \mathrm{Frac}(R)$. If A is an R-module, prove that every element in $Q \otimes_R A$ has the form $q \otimes a$ for $q \in Q$ and $a \in A$ (instead of $\sum_i q_i \otimes a_i$). (Compare this result with Example 8.89.)

8.48 Let m and n be positive integers, and let $d = (m, n)$. Prove that there is an isomorphism of abelian groups

$$\mathbb{I}_m \otimes \mathbb{I}_n \cong \mathbb{I}_d.$$

8.49 Let k be a commutative ring, and let P and Q be projective k-modules. Prove that $P \otimes_k Q$ is a projective k-module.

8.50 Let k be a commutative ring, and let P and Q be flat k-modules. Prove that $P \otimes_k Q$ is a flat k-module.

8.51 Assume that the following diagram commutes, and that the vertical arrows are isomorphisms.

$$\begin{array}{ccccccccc} 0 & \longrightarrow & A' & \longrightarrow & A & \longrightarrow & A'' & \longrightarrow & 0 \\ & & \downarrow & & \downarrow & & \downarrow & & \\ 0 & \longrightarrow & B' & \longrightarrow & B & \longrightarrow & B'' & \longrightarrow & 0 \end{array}$$

Prove that the bottom row is exact if and only if the top row is exact.

8.52 ***(Five Lemma).*** Consider a commutative diagram with exact rows

$$\begin{array}{ccccccccc} A_1 & \longrightarrow & A_2 & \longrightarrow & A_3 & \longrightarrow & A_4 & \longrightarrow & A_5 \\ {\scriptstyle h_1}\downarrow & & {\scriptstyle h_2}\downarrow & & \downarrow{\scriptstyle h_3} & & \downarrow{\scriptstyle h_4} & & \downarrow{\scriptstyle h_5} \\ B_1 & \longrightarrow & B_2 & \longrightarrow & B_3 & \longrightarrow & B_4 & \longrightarrow & B_5 \end{array}$$

(i) If h_2 and h_4 are surjective and h_5 is injective, prove that h_3 is surjective.

(ii) If h_2 and h_4 are injective and h_1 is surjective, prove that h_3 is injective.

(iii) If h_1, h_2, h_4, and h_5 are isomorphisms, prove that h_3 is an isomorphism.

8.53 Prove that a ring R is left noetherian if and only if every direct limit (with directed index set) of injective left R-modules is itself injective.
Hint. See Theorem 8.105.

8.54 Let $\mathcal{A}$, $\mathcal{B}$, and $\mathcal{C}$ be categories. A ***functor of two variables*** $T : \mathcal{A} \times \mathcal{B} \to \mathcal{C}$ assigns, to each ordered pair of objects (A, B), where $A \in \text{ob}(\mathcal{A})$ and $B \in \text{ob}(\mathcal{B})$, an object $T(A, B) \in \text{ob}(\mathcal{C})$, and to each ordered pair of morphisms $f : A \to A'$ in $\mathcal{A}$ and $g : B \to B'$ in $\mathcal{B}$, a morphism $T(f, g) : T(A, B) \to T(A', B')$, such that

(a) Fixing either variable is a functor: for example, if $A \in \text{ob}(\mathcal{A})$, then

$$T_A = T(A, \quad) : \mathcal{B} \to \mathcal{C}$$

is a functor, where $T_A(B) = T(A, B)$ and $T_A(g) = T(1_A, g)$.

(b) The following diagram commutes:

$$\begin{array}{ccc} T(A, B) & \xrightarrow{T(1_A, g)} & T(A, B') \\ {\scriptstyle T(f, 1_B)}\downarrow & \searrow^{T(f,g)} & \downarrow{\scriptstyle T(f, 1_{B'})} \\ T(A', B) & \xrightarrow[T(1_{A'}, g)]{} & T(A', B') \end{array}$$

(i) Prove that $\otimes : \textbf{Mod}_R \times {}_R\textbf{Mod} \to \textbf{Ab}$ is a functor of two variables.

(ii) Modify the definition of a functor of two variables to allow contravariance in a variable, and prove that Hom is a functor of two variables.

8.5 CHARACTERS

Representation theory is the study of homomorphisms of abstract groups G into groups of nonsingular matrices; such homomorphisms produce numerical invariants whose arithmetic properties help to prove theorems about G. We now introduce this vast subject with one goal being a proof of the following theorem.

Theorem 8.115 (Burnside). *Every group G of order $p^m q^n$, where p and q are primes, is a solvable group.*

Notice that Burnside's theorem cannot be improved to groups having orders with only three distinct prime factors, for A_5 is a simple group of order $60 = 2^2 \cdot 3 \cdot 5$.

Using representations, we will prove the following theorem.

Theorem. *If G is a nonabelian finite simple group, then $\{1\}$ is the only conjugacy class whose size is a prime power.*

Proposition 8.116. *The preceding theorem implies Burnside's theorem.*

Proof. Assume that Burnside's theorem is false, and let G be a "least criminal"; that is, G is a counterexample of smallest order. If G has a proper normal subgroup H with $H \neq \{1\}$, then both H and G/H are solvable, for their orders are smaller than $|G|$ and are of the form $p^i q^j$. By Proposition 4.24, G is solvable, and this is a contradiction. We may assume, therefore, that G is a nonabelian simple group.

Let Q be a Sylow q-subgroup of G. If $Q = \{1\}$, then G is a p-group, contradicting G being a nonabelian simple group; hence, $Q \neq \{1\}$. Since the center of Q is nontrivial, by Theorem 2.103, we may choose a nontrivial element $x \in Z(Q)$. Now $Q \leq C_G(x)$, for every element in Q commutes with x, and so

$$[G : Q] = [G : C_G(x)][C_G(x) : Q];$$

that is, $[G : C_G(x)]$ is a divisor of $[G : Q] = p^m$. Of course, $[G : C_G(x)]$ is the number of elements in the conjugacy class x^G of x (Corollary 2.100), and so the hypothesis says that $|x^G| = 1$; hence, $x \in Z(G)$, which contradicts G being simple. •

The proof that the hypothesis of the proposition is true will use representation theory (see Theorem 8.153).

We now specialize the definition of k-representation on page 550 from arbitrary fields k of scalars to the complex numbers $\mathbb{C}$.

Definition. A ***representation*** of a group G is a homomorphism

$$\sigma : G \to \mathrm{GL}(V),$$

where V is a vector space over $\mathbb{C}$. The ***degree*** of σ is $\dim(V)$.

For the remainder of this section, we restrict ourselves to finite groups and representations having finite degree. If a representation $\sigma : G \to \mathrm{GL}(V)$ has degree n and one chooses a basis of V, then each $\sigma(g)$ can be regarded as an $n \times n$ nonsingular matrix with entries in $\mathbb{C}$.

Representations can be translated into the language of modules. In Proposition 8.37, we proved that every representation $\sigma : G \to \mathrm{GL}(V)$ equips V with the structure of a left $\mathbb{C}G$-module (and conversely): If $g \in G$, then $\sigma(g) : V \to V$, and we define scalar multiplication gv, for $g \in G$ and $v \in V$, by

$$gv = \sigma(g)(v).$$

Example 8.117.
We now show that permutation representations, that is, G-sets,[9] give a special kind of representation. A G-set X corresponds to a homomorphism $\pi : G \to S_X$, where S_X is the symmetric group of all permutations of X. If V is the complex vector space having X as a basis, then we may regard $S_X \leq \mathrm{GL}(V)$ in the following way. Each permutation

[9]Recall that if a group G acts on a set X, then X is called a *G-set*.

$\pi(g)$ of X, where $g \in G$, is now a permutation of a basis of V and, hence, it determines a nonsingular linear transformation on V. With respect to the basis X, the matrix of $\pi(g)$ is a ***permutation matrix***: It arises by permuting the columns of the identity matrix I by $\pi(g)$; thus, it has exactly one entry equal to 1 in each row and column while all its other entries are 0. ◂

One of the most important representations is the *regular representation*; in terms of modules, the regular representation is the group algebra $\mathbb{C}G$ regarded as a left module over itself.

Definition. If G is a group, then the representation $\rho\colon G \to \mathrm{GL}(\mathbb{C}G)$ defined, for all $g, h \in G$, by

$$\rho(g)\colon h \mapsto gh,$$

is called the ***regular representation***.

Two representations $\sigma\colon G \to \mathrm{GL}(V)$ and $\tau\colon G \to \mathrm{GL}(W)$ can be added.

Definition. If $\sigma\colon G \to \mathrm{GL}(V)$ and $\tau\colon G \to \mathrm{GL}(W)$ are representations, then their ***sum*** $\sigma + \tau\colon G \to \mathrm{GL}(V \oplus W)$ is defined by

$$(\sigma + \tau)(g)\colon (v, w) \mapsto (\sigma(g)v, \tau(g)w)$$

for all $g \in G$, $v \in V$, and $w \in W$.

In matrix terms, if $\sigma\colon G \to \mathrm{GL}(n, \mathbb{C})$ and $\tau\colon G \to \mathrm{GL}(m, \mathbb{C})$, then

$$\sigma + \tau\colon G \to \mathrm{GL}(n + m, \mathbb{C}),$$

and if $g \in G$, then $(\sigma + \tau)(g)$ is the direct sum of blocks $\sigma(g) \oplus \tau(g)$; that is,

$$(\sigma + \tau)(g) = \begin{bmatrix} \sigma(g) & 0 \\ 0 & \tau(g) \end{bmatrix}.$$

The following terminology is the common one used in group representations.

Definition. A representation σ of a group G is ***irrreducible*** if the corresponding $\mathbb{C}G$-module is simple; a representation σ is ***completely reducible*** if it is a direct sum of irreducible representations; that is, the corresponding $\mathbb{C}G$-module is semisimple.

Example 8.118.
A representation σ is ***linear*** if $\mathrm{degree}(\sigma) = 1$. The trivial representation of any group G is linear, for the principal module $V_0(\mathbb{C})$ is one-dimensional. If $G = S_n$, then $\mathrm{sgn}\colon G \to \{\pm 1\}$ is also a linear representation.

Every linear representation is irreducible, for the corresponding $\mathbb{C}G$-module must be simple; after all, every submodule is a subspace, and $\{0\}$ and V are the only subspaces of a one-dimensional vector space V. It follows that the trivial representation of any group G is irreducible, as is the representation sgn of S_n. ◂

Recall the proof of the Wedderburn–Artin theorem: There are pairwise nonisomorphic minimal left ideals $L_1, \ldots, L_r$ in $\mathbb{C}G$ and $\mathbb{C}G = B_1 \oplus \cdots \oplus B_r$, where B_i is generated by all minimal left ideals isomorphic to L_i. Now $B_i \cong \mathrm{Mat}_{n_i}(\mathbb{C})$, by Corollary 8.65. But all minimal left ideals in $\mathrm{Mat}_{n_i}(\mathbb{C})$ are isomorphic, by Lemma 8.61(ii), so that $L_i \cong \mathrm{COL}(1) \cong \mathbb{C}^{n_i}$ (see Example 8.30). Therefore,

$$B_i \cong \mathrm{End}(L_i),$$

where we have abbreviated $\mathrm{End}_{\mathbb{C}}(L_i)$ to $\mathrm{End}(L_i)$.

Proposition 8.119.

(i) *For each minimal left ideal L_i in $\mathbb{C}G$, there is an irreducible representation $\lambda_i : G \to \mathrm{GL}(L_i)$, given by left multiplication*:

$$\lambda_i(g) : u_i \mapsto g u_i,$$

where $g \in G$ and $u_i \in L_i$; moreover, $\mathrm{degree}(\lambda_i) = n_i = \dim(L_i)$.

(ii) *The representation λ_i extends to a $\mathbb{C}$-algebra map $\widetilde{\lambda}_i : \mathbb{C}G \to \mathbb{C}G$ if we define*

$$\widetilde{\lambda}_i(g) u_j = \begin{cases} g u_i & if\ j = i \\ 0 & if\ j \neq i \end{cases} \tag{2}$$

for $g \in G$ and $u_j \in B_j$.

Proof. (i) Since L_i is a left ideal in $\mathbb{C}G$, each $g \in G$ acts on L_i by left multiplication, and so the corresponding representation λ_i of G is as stated; it is an irreducible representation because minimal left ideals are simple modules.

(ii) If we regard $\mathbb{C}G$ and $\mathrm{End}(L_i)$ as vector spaces over $\mathbb{C}$, then λ_i extends to a linear transformation $\widetilde{\lambda}_i : \mathbb{C}G \to \mathrm{End}(L_i)$ (because the elements of G are a basis of $\mathbb{C}G$):

$$\widetilde{\lambda}_i : \sum_g c_g g \mapsto \sum_g c_g \lambda_i(g).$$

Let us show that $\widetilde{\lambda}_i : \mathbb{C}G \to \mathrm{End}(L_i)$ is actually a $\mathbb{C}$-algebra map. If $u_i \in L_i$ and $g, h \in G$, then

$$\widetilde{\lambda}_i(gh) : u_i \mapsto (gh) u_i,$$

while

$$\widetilde{\lambda}_i(g)\widetilde{\lambda}_i(h) : u_i \mapsto h u_i \mapsto g(h u_i);$$

these are the same, by associativity.

At the moment, $\widetilde{\lambda}_i(g) u_i$ is defined only for $u_i \in B_i = \mathrm{End}(L_i)$. For each $g \in G$, we now extend the map $\widetilde{\lambda}_i(g)$ to $\mathbb{C}G = B_1 \oplus \cdots \oplus B_r$ by defining $\widetilde{\lambda}_i(g) u_j = 0$, where $u_j \in B_j \cong \mathrm{End}(L_j)$ and $j \neq i$. The extended map $\widetilde{\lambda}_i(g)$ (we keep the same notation even though its target has been enlarged from B_i to $\mathbb{C}G$) is also a $\mathbb{C}G$-algebra map. If $j \neq i$, then $u_i u_j \in B_i B_j = \{0\}$, so that $\widetilde{\lambda}_i(g)(u_i u_j) = 0$; on the other hand, $(\widetilde{\lambda}_i(g) u_i)(\widetilde{\lambda}_i(g) u_j) = 0$, by definition. •

It is natural to call two representations *equivalent* if their corresponding modules are isomorphic.

Definition. If G is a group and $\sigma, \tau \colon G \to \mathrm{GL}(n, \mathbb{C})$ are representations, then σ and τ are ***equivalent***, denoted by $\sigma \sim \tau$, if there is a nonsingular $n \times n$ matrix P that intertwines them; that is,

$$P\sigma(g)P^{-1} = \tau(g)$$

for every $g \in G$.

Of course, this definition comes from Corollary 8.39, which says that the $\mathbb{C}G$-modules $(\mathbb{C}^n)^\sigma$ and $(\mathbb{C}^n)^\tau$ are isomorphic as $\mathbb{C}G$-modules if and only if $\sigma \sim \tau$.

Corollary 8.120.

(i) *Every irreducible representation of a finite group G is equivalent to one of the representations λ_i given in* Proposition 8.119(i).

(ii) *Every irreducible representation of a finite abelian group is linear.*

(iii) *If $\sigma \colon G \to \mathrm{GL}(V)$ is a representation of a finite group G, then $\sigma(g)$ is similar to a diagonal matrix for each $g \in G$.*

Proof. (i) If $\sigma \colon G \to \mathrm{GL}(V)$ is an irreducible representation σ, then the corresponding $\mathbb{C}G$-module V^σ is a simple module. Therefore, $V^\sigma \cong L_i$, for some i, by Proposition 8.54. But $L_i \cong V^{\lambda_i}$, so that $V^\sigma \cong V^{\lambda_i}$ and $\sigma \sim \lambda_i$.

(ii) Since G is abelian, $\mathbb{C}G = \sum_i B_i$ is commutative, and so all $n_i = 1$. But $n_i = \mathrm{degree}(\lambda_i)$.

(iii) If $\sigma' = \sigma|\langle g\rangle$, then $\sigma'(g) = \sigma(g)$. Now σ' is a representation of the abelian group $\langle g\rangle$, and so part (ii) implies that the module $V^{\langle g\rangle}$ is a direct sum of one-dimensional submodules. If $V^{\langle g\rangle} = \langle v_1\rangle \oplus \cdots \oplus \langle v_m\rangle$, then the matrix of $\sigma(g)$ with respect to the basis $v_1, \ldots, v_m$ is diagonal. •

Example 8.121.

(i) The Wedderburn–Artin theorem can be restated to say that every representation $\tau \colon G \to \mathrm{GL}(V)$ is completely reducible: $\tau = \sigma_1 + \cdots + \sigma_k$, where each σ_j is irreducible; moreover, the multiplicity of each σ_j is uniquely determined by τ. Since each σ_j is equivalent to some λ_i, we usually collect terms and write $\tau \sim \sum_i m_i\lambda_i$, where the multiplicities m_i are nonnegative integers.

(ii) The regular representation $\rho \colon G \to \mathbb{C}G$ is important because every irreducible representation is a summand of it. Now ρ is equivalent to the sum

$$\rho \sim n_1\lambda_1 + \cdots + n_r\lambda_r,$$

where n_i is the degree of λ_i [remember that $\mathbb{C}G = \sum_i B_i$, where $B_i \cong \mathrm{End}(L_i) \cong \mathrm{Mat}_{n_i}(\mathbb{C})$; as a $\mathbb{C}G$-module, the simple module L_i can be viewed as the first columns of $n_i \times n_i$ matrices, and so B_i is a direct sum of n_i copies of L_i]. ◂

Recall that the ***trace*** of an $n \times n$ matrix $A = [a_{ij}]$ with entries in a commutative ring k is the sum of the diagonal entries: $\operatorname{tr}(A) = \sum_{i=1}^{n} a_{ii}$.

When k is a field, then $\operatorname{tr}(A)$ turns out to be the sum of the eigenvalues of A (we will assume this result now, but it is more convenient for us to prove it in the next chapter). Here are two other elementary facts about the trace that we will prove now.

Proposition 8.122.

(i) *If $A = [a_{ij}]$ and $B = [b_{ij}]$ are $n \times n$ matrices with entries in a commutative ring k, then*

$$\operatorname{tr}(A + B) = \operatorname{tr}(A) + \operatorname{tr}(B) \qquad \text{and} \qquad \operatorname{tr}(AB) = \operatorname{tr}(BA).$$

(ii) *If $B = PAP^{-1}$, then* $\operatorname{tr}(B) = \operatorname{tr}(A)$.

Proof. (i) The additivity of trace follows from the diagonal entries of $A+B$ being $a_{ii}+b_{ii}$. If $(AB)_{ii}$ denotes the ii entry of AB, then

$$(AB)_{ii} = \sum_j a_{ij} b_{ji},$$

and so

$$\operatorname{tr}(AB) = \sum_i (AB)_{ii} = \sum_{i,j} a_{ij} b_{ji}.$$

Similarly,

$$\operatorname{tr}(BA) = \sum_{j,i} b_{ji} a_{ij}.$$

The entries commute because they lie in the commutative ring k, and so $a_{ij} b_{ji} = b_{ji} a_{ij}$ for all i, j. It follows that $\operatorname{tr}(AB) = \operatorname{tr}(BA)$, as desired.

(ii)

$$\operatorname{tr}(B) = \operatorname{tr}\big((PA)P^{-1}\big) = \operatorname{tr}\big(P^{-1}(PA)\big) = \operatorname{tr}(A). \quad \bullet$$

It follows from (ii) that we can define the trace of a linear transformation $T : V \to V$, where V is a vector space over a field k, as the trace of any matrix arising from it: If A and B are matrices of T, determined by two choices of bases of V, then $B = PAP^{-1}$ for some nonsingular matrix P, and so $\operatorname{tr}(B) = \operatorname{tr}(A)$.

Definition. If $\sigma : G \to \mathrm{GL}(V)$ is a representation, then its ***character*** is the function $\chi_\sigma : G \to \mathbb{C}$ defined by

$$\chi_\sigma(g) = \operatorname{tr}(\sigma(g));$$

we call χ_σ the character ***afforded*** by σ. An ***irreducible character*** is a character afforded by an irreducible representation. The ***degree*** of χ_σ is defined to be the degree of σ; that is,

$$\operatorname{degree}(\chi_\sigma) = \operatorname{degree}(\sigma) = \dim(V).$$

Example 8.123.

(i) The character θ afforded by a linear representation (see Example 8.118) is called a ***linear character***; that is, $\theta = \chi_\sigma$, where degree$(\sigma) = 1$. Since every linear representation is simple, every linear character is irreducible.

(ii) The representation $\lambda_i : G \to \mathrm{GL}(L_i)$ [see Proposition 8.119(i)] is irreducible. Thus, the character

$$\chi_i = \chi_{\lambda_i}$$

afforded by λ_i is irreducible.

(iii) In light of Proposition 8.119(ii), it makes sense to speak of $\chi_i(u)$ for every $u \in \mathbb{C}G$. Of course, $\chi_i(u_j) = 0$ for all $u_j \in \mathrm{End}(L_j)$ when $j \neq i$, so that

$$\chi_i(u_j) = \begin{cases} \mathrm{tr}(\widetilde{\lambda}_i(u_j)) & \text{if } j = i \\ 0 & \text{if } j \neq i. \end{cases}$$

(iv) If $\sigma : G \to \mathrm{GL}(V)$ is any representation, then $\chi_\sigma(1) = n$, where n is the degree of σ. After all, $\sigma(1)$ is the identity matrix, and its trace is $n = \dim(V)$.

(v) Let $\sigma : G \to S_X$ be a homomorphism; as in Example 8.117, we may regard σ as a representation on V, where V is the vector space over $\mathbb{C}$ with basis X. For every $g \in G$, the matrix $\sigma(g)$ is a permutation matrix, and its xth diagonal entry is 1 if $\sigma(g)x = x$; otherwise, it is 0. Thus,

$$\chi_\sigma(g) = \mathrm{tr}(\sigma(g)) = \mathrm{Fix}(\sigma(g)),$$

the number of $x \in X$ fixed by $\sigma(g)$. In other words, if X is a G-set, then we may view each $g \in G$ as acting on X, and the number of ***fixed points*** of the action of g is a character value (see Example 8.144 for a related discussion). ◀

Characters are compatible with addition of representations: If $\sigma : G \to \mathrm{GL}(V)$ and $\tau : G \to \mathrm{GL}(W)$, then $\sigma + \tau : G \to \mathrm{GL}(V \oplus W)$, and

$$\mathrm{tr}((\sigma + \tau)(g)) = \mathrm{tr}\left(\begin{bmatrix} \sigma(g) & 0 \\ 0 & \tau(g) \end{bmatrix}\right) = \mathrm{tr}(\sigma(g)) + \mathrm{tr}(\tau(g)).$$

Therefore,

$$\chi_{\sigma+\tau} = \chi_\sigma + \chi_\tau.$$

If σ and τ are equivalent representations, then

$$\mathrm{tr}(\sigma(g)) = \mathrm{tr}(P\sigma(g)P^{-1}) = \mathrm{tr}(\tau(g))$$

for all $g \in G$; that is, they have the same characters: $\chi_\sigma = \chi_\tau$. It follows that if $\sigma : G \to \mathrm{GL}(V)$ is a representation, then its character χ_σ can be computed relative to any convenient basis of V.

Proposition 8.124.

(i) *Every character χ_σ is an $\mathbb{N}$-linear combination of the irreducible characters $\chi_i = \chi_{\lambda_i}$ afforded by $\lambda_i : G \to \mathrm{GL}(L_i)$: there are integers $m_i \geq 0$ with*

$$\chi_\sigma = \sum_i m_i \chi_i.$$

(ii) *Equivalent representations have the same character.*

(iii) *The only irreducible characters of G are $\chi_1, \ldots, \chi_r$.*

Proof. (i) The character χ_σ arises from a representation σ of G, which, in turn, arises from a $\mathbb{C}G$-module V. But V is a semisimple module (because $\mathbb{C}G$ is a semisimple ring), and so V is a direct sum of simple modules: $V = \sum_j S_j$. By Proposition 8.54, each $S_j \cong L_i$ for some minimal left ideal L_i. If, for each i, we let $m_i \geq 0$ be the number of S_j isomorphic to L_i, then $\chi_\sigma = \sum_i m_i \chi_i$.

(ii) This follows from part (ii) of Proposition 8.122 and Corollary 8.120(i).

(iii) This follows from part (ii) and Corollary 8.120(i). •

As a consequence of the proposition, we call $\chi_1, \ldots, \chi_r$ ***the irreducible characters*** of G.

Example 8.125.
(i) The (linear) character χ_1 afforded by the trivial representation $\sigma : G \to \mathbb{C}$ with $\sigma(g) = 1$ for all $g \in G$ is called the ***trivial character***. Thus, $\chi_1(g) = 1$ for all $g \in G$.

(ii) Let us compute the ***regular character*** $\psi = \chi_\rho$ afforded by the regular representation $\rho : G \to \mathrm{GL}(\mathbb{C}G)$, where $\rho(g) : u \mapsto gu$ for all $g \in G$ and $u \in \mathbb{C}G$. Any basis of $\mathbb{C}G$ can be used for this computation; we choose the usual basis comprised of the elements of G. If $g = 1$, then Example 8.123(iv) shows that $\psi(1) = \dim(\mathbb{C}G) = |G|$. On the other hand, if $g \neq 1$, then for all $h \in G$, we have gh a basis element distinct from h. Therefore, the matrix of $\rho(g)$ has 0's on the diagonal, and so its trace is 0. Thus,

$$\psi(g) = \begin{cases} 0 & \text{if } g \neq 1 \\ |G| & \text{if } g = 1. \end{cases} \quad ◀$$

We have already proved that equivalent representations have the same character. The coming discussion will give the converse: If two representations have the same character, then they are equivalent.

Definition. A function $\varphi : G \to \mathbb{C}$ is a ***class function*** if it is constant on conjugacy classes; that is, if $h = xgx^{-1}$, then $\varphi(h) = \varphi(g)$.

Every character χ_σ afforded by a representation σ is a class function: If $h = xgx^{-1}$, then

$$\sigma(h) = \sigma(xgx^{-1}) = \sigma(x)\sigma(g)\sigma(x)^{-1},$$

and so $\mathrm{tr}(\sigma(h)) = \mathrm{tr}(\sigma(g))$; that is,

$$\chi_\sigma(h) = \chi_\sigma(g).$$

Not every class function is a character. For example, if χ is a character, then $-\chi$ is a class function; it is not a character because $-\chi(1)$ is negative, and so it cannot be a degree.

Definition. We denote the set of all class functions $G \to \mathbb{C}$ by $\mathrm{cf}(G)$:

$$\mathrm{cf}(G) = \{\varphi\colon G \to \mathbb{C} : \varphi(g) = \varphi(xgx^{-1}) \text{ for all } x, g \in G\}.$$

It is easy to see that $\mathrm{cf}(G)$ is a vector space over $\mathbb{C}$.

An element $u = \sum_{g \in G} c_g g \in \mathbb{C}G$ is an n-tuple (c_g) of complex numbers; that is, u is a function $u\colon G \to \mathbb{C}$ with $u(g) = c_g$ for all $g \in G$. From this viewpoint, we see that $\mathrm{cf}(G)$ is a subring of $\mathbb{C}G$. Note that a class function is a class sum; therefore, Lemma 8.68 says that $\mathrm{cf}(G)$ is the center $Z(\mathbb{C}G)$, and so

$$\dim(\mathrm{cf}(G)) = r,$$

where r is the number of conjugacy classes in G (see Theorem 8.69).

Definition. Write $\mathbb{C}G = B_1 \oplus \cdots \oplus B_r$, where $B_i \cong \mathrm{End}(L_i)$, and let e_i denote the identity element of B_i; hence,

$$1 = e_1 + \cdots e_r,$$

where 1 is the identity element of $\mathbb{C}G$. The elements e_i are called the ***idempotents*** in $\mathbb{C}G$.

Not only is each e_i an idempotent, that is, $e_i^2 = e_i$, but it is easy to see that

$$e_i e_j = \delta_{ij} e_i,$$

where δ_{ij} is the Kronecker delta.

Lemma 8.126. *The irreducible characters* $\chi_1, \ldots, \chi_r$ *form a basis of* $\mathrm{cf}(G)$.

Proof. We have just seen that $\dim(\mathrm{cf}(G)) = r$, and so it suffices to prove that $\chi_1, \ldots, \chi_r$ is a linearly independent list, by Corollary 3.89(ii). We have already noted that $\chi_i(u_j) = 0$ for all $j \neq i$; in particular, $\chi_i(e_j) = 0$. On the other hand, $\chi_i(e_i) = n_i$, where n_i is the degree of χ_i, for it is the trace of the $n_i \times n_i$ identity matrix.

Suppose now that $\sum_i c_i \chi_i = 0$. It follows, for all j, that

$$0 = \Big(\sum_i c_i \chi_i\Big)(e_j) = c_j \chi_j(e_j) = c_j n_j.$$

Therefore, all $c_j = 0$, as desired. •

Theorem 8.127. *Two representations of a finite group G are equivalent if and only if they afford the same character: $\chi_\sigma = \chi_\tau$.*

Proof. We have already proved necessity, in Proposition 8.124(ii). For sufficiency, Proposition 8.124(ii) says that every representation is completely reducible: There are nonnegative integers m_i and ℓ_i with $\sigma \sim \sum_i m_i\lambda_i$ and $\tau \sim \sum_i \ell_i\lambda_i$. By hypothesis, the corresponding characters coincide:

$$\sum_i m_i\chi_i = \chi_\sigma = \chi_\tau = \sum_i \ell_i\chi_i.$$

As the irreducible characters $\chi_1, \ldots, \chi_r$ are a basis of $\mathrm{cf}(G)$, $m_i = \ell_i$ for all i, and so $\sigma \sim \tau$. •

There are relations between the irreducible characters that facilitate their calculation. We begin by finding the expression of the idempotents e_i in terms of the basis G of $\mathbb{C}G$. By Example 8.123(iv), $\chi_i(1) = n_i$, the degree of λ_i. On the other hand, by Eq. (2) in Proposition 8.119, we have $\chi_i(e_j) = 0$ if $j \neq i$, so that

$$n_i = \chi_i(1) = \sum_j \chi_i(e_j) = \chi_i(e_i). \tag{3}$$

We also observe, for all $y \in G$, that

$$\chi_i(e_i y) = \chi_i(y), \tag{4}$$

for $y = \sum_j e_j y$, and so $\chi_i(y) = \sum_j \chi_i(e_j y) = \chi_i(e_i y)$, because $e_j y \in B_j$.

Proposition 8.128. *If $e_i = \sum_{g\in G} a_{ig} g$, where $a_{ig} \in \mathbb{C}$, then*

$$a_{ig} = \frac{n_i\chi_i(g^{-1})}{|G|}.$$

Proof. Let ψ be the regular character; that is, ψ is the character afforded by the regular representation. Now $e_i g^{-1} = \sum_h a_{ih} hg^{-1}$, so that

$$\psi(e_i g^{-1}) = \sum_{h\in G} a_{ih}\psi(hg^{-1}).$$

By Example 8.125(ii), $\psi(1) = |G|$ when $h = g$ and $\psi(hg^{-1}) = 0$ when $h \neq g$. Therefore,

$$a_{ig} = \frac{\psi(e_i g^{-1})}{|G|}.$$

On the other hand, since $\psi = \sum_j n_j\chi_j$, we have

$$\psi(e_i g^{-1}) = \sum_j n_j\chi_j(e_i g^{-1}) = n_i\chi_i(e_i g^{-1}),$$

by Eq. (2) in Proposition 8.119. But $\chi_i(e_i g^{-1}) = \chi_i(g^{-1})$, by Eq. (4). Therefore, $a_{ig} = n_i\chi_i(g^{-1})/|G|$. •

It is now convenient to equip $\mathrm{cf}(G)$ with an inner product.

Definition. If $\alpha, \beta \in \mathrm{cf}(G)$, define

$$(\alpha, \beta) = \frac{1}{|G|} \sum_{g \in G} \alpha(g) \overline{\beta(g)},$$

where $\overline{c}$ denotes the complex conjugate of a complex number c.

It is easy to see that we have defined an inner product;[10] that is, for all $c_1, c_2 \in \mathbb{C}$,

(i) $(c_1\alpha_1 + c_2\alpha_2, \beta) = c_1(\alpha_1, \beta) + c_2(\alpha_2, \beta)$;

(ii) $(\beta, \alpha) = \overline{(\alpha, \beta)}$.

Note that (α, α) is real, by (ii), and the inner product is *definite*; that is, $(\alpha, \alpha) > 0$ if $\alpha \neq 0$.

Theorem 8.129. *With respect to the inner product just defined, the irreducible characters $\chi_1, \ldots, \chi_r$ form an orthonormal basis; that is,*

$$(\chi_i, \chi_j) = \delta_{ij}.$$

Proof. By Proposition 8.128, we have

$$e_j = \frac{1}{|G|} \sum_g n_j \chi_j(g^{-1}) g.$$

Hence,

$$\begin{aligned} \chi_i(e_j)/n_j &= \frac{1}{|G|} \sum_g \chi_j(g^{-1}) \chi_i(g) \\ &= \frac{1}{|G|} \sum_g \chi_i(g) \overline{\chi_j(g)} \\ &= (\chi_i, \chi_j); \end{aligned}$$

the next to last equation follows from Exercise 8.56(ii) on page 632, for χ_j is a character (not merely a class function), and so $\chi_j(g^{-1}) = \overline{\chi_j(g)}$. The result now follows, for $\chi_i(e_j)/n_j = \delta_{ij}$, by Eqs. (2) and (3). •

The inner product on $\mathrm{cf}(G)$ can be used to check irreducibility.

Definition. A ***generalized character*** φ on a finite group G is a linear combination

$$\varphi = \sum_i m_i \chi_i,$$

where $\chi_1, \ldots, \chi_r$ are the irreducible characters of G and the coefficients $m_i \in \mathbb{Z}$.

If θ is a character, then $\theta = \sum_i m_i \chi_i$, where all the coefficients are *nonnegative* integers, by Proposition 8.124.

[10] This inner product is *not* a bilinear form because we have $(\beta, \alpha) = \overline{(\alpha, \beta)}$, not $(\beta, \alpha) = (\alpha, \beta)$. Such a function is often called a *Hermitian form* or a *sesquilinear form* (*sesqui* means "one and a half").

Corollary 8.130. *A generalized character θ of a group G is an irreducible character if and only if $\theta(1) > 0$ and*

$$(\theta, \theta) = 1.$$

Proof. If θ is an irreducible character, then $\theta = \chi_i$ for some i, and so $(\theta, \theta) = (\chi_i, \chi_i) = 1$. Moreover, $\theta(1) = \deg(\chi_i) > 0$.

Conversely, let $\theta = \sum_j m_j \chi_j$, where $m_j \in \mathbb{Z}$, and suppose that $(\theta, \theta) = 1$. Then $1 = \sum_j m_j^2$; hence, some $m_i^2 = 1$ and all other $m_j = 0$. Therefore, $\theta = \pm\chi_i$, and so $\theta(1) = \pm\chi_i(1)$. Since $\chi_i(1) = \deg(\chi_i) > 0$, the hypothesis $\theta(1) > 0$ gives $m_i = 1$. Therefore, $\theta = \chi_i$, and so θ is an irreducible character. •

Let us assemble the notation we will use from now on.

Notation. If G is a finite group, we denote its conjugacy classes by

$$C_1, \ldots, C_r,$$

a choice of elements, one from each conjugacy class, by

$$g_1 \in C_1, \ldots, g_r \in C_r,$$

its irreducible characters by

$$\chi_1, \ldots, \chi_r,$$

their degrees by

$$n_1 = \chi_1(1), \ldots, n_r = \chi_r(1),$$

and the sizes of the conjugacy classes by

$$h_1 = |C_1|, \ldots, h_r = |C_r|.$$

The matrix $[\chi_i(g_j)]$ is a useful way to display information.

Definition. The ***character table*** of G is the $r \times r$ complex matrix whose ij entry is $\chi_i(g_j)$.

We always assume that $C_1 = \{1\}$ and that χ_1 is the trivial character. Thus, the first row consists of all 1's, while the first column consists of the degrees of the characters: $\chi_i(1) = n_i$ for all i, by Example 8.123(iv). The ith row of the character table consists of the values

$$\chi_i(1), \chi_i(g_2), \ldots, \chi_i(g_r).$$

There is no obvious way of labeling the other conjugacy classes (or the other irreducible characters), so that a finite group G has many character tables. Nevertheless, we usually speak of "the" character table of G.

Since the inner product on $\text{cf}(G)$ is summed over all $g \in G$, not just the chosen g_i (one from each conjugacy class), it can be rewritten as a "weighted" inner product:

$$(\chi_i, \chi_j) = \frac{1}{|G|} \sum_{k=1}^{r} h_k \chi_i(g_k) \overline{\chi_j(g_k)}.$$

Theorem 8.129 says that the weighted inner product of distinct rows in the character table is 0, while the weighted inner product of any row with itself is 1.

Example 8.131.
A character table can have complex entries. For example, it is easy to see that the character table for a cyclic group $G = \langle x \rangle$ of order 3 is given in Table 8.1, where $\omega = e^{2\pi i/3}$ is a primitive cube root of unity.

g_i	1	x	x^2
h_i	1	1	1
χ_1	1	1	1
χ_2	1	ω	ω^2
χ_3	1	ω^2	ω

Table 8.1. Character Table of $\mathbb{I}_3$ ◄

Example 8.132.
Write the four-group in additive notation:

$$\mathbf{V} = \{0, a, b, a+b\}.$$

As a vector space over $\mathbb{F}_2$, $\mathbf{V}$ has basis a, b, and the "coordinate functions" on $\mathbf{V}$, which take values in $\{1, -1\} \subseteq \mathbb{C}$, are linear, hence irreducible, representations. For example, the character χ_2 arising from the function that is nontrivial on a and trivial on b is

$$\chi_2(v) = \begin{cases} -1 & \text{if } v = a \quad \text{or} \quad v = a+b \\ 1 & \text{if } v = 0 \quad \text{or} \quad v = b. \end{cases}$$

Table 8.2 is the character table.

g_i	0	a	b	$a+b$
h_i	1	1	1	1
χ_1	1	1	1	1
χ_2	1	-1	1	-1
χ_3	1	1	-1	-1
χ_4	1	-1	-1	1

Table 8.2. Character Table of $\mathbf{V}$ ◄

Example 8.133.
Table 8.3 on page 618 is the character table for the symmetric group $G = S_3$. Since two permutations in S_n are conjugate if and only if they have the same cycle structure, there are three conjugacy classes, and we choose elements 1, (1 2), and (1 2 3) from each. In Example 8.71(i), we saw that there are three irreducible representations: $\lambda_1 =$ the trivial representation, $\lambda_2 =$ sgn, and a third representation λ_3 of degree 2. We now give the character table, after which we discuss its entries.

g_i	1	(1 2)	(1 2 3)
h_i	1	3	2
χ_1	1	1	1
χ_2	1	−1	1
χ_3	2	0	−1

Table 8.3. Character Table of S_3

We have already discussed the first row and column of any character table. Since $\chi_2 = \text{sgn}$, the second row records the fact that 1 and (1 2 3) are even while (1 2) is odd. The third row has entries

$$2 \quad a \quad b,$$

where a and b are to be found. The weighted inner products of row 3 with the other two rows gives the equations

$$2 + 3a + 2b = 0$$
$$2 - 3a + 2b = 0.$$

It follows easily that $a = 0$ and $b = -1$. ◂

The following lemma will be used to describe the inner products of the columns of the character table.

Lemma 8.134. *If A is the character table of a finite group G, then A is nonsingular and its inverse A^{-1} has ij entry*

$$(A^{-1})_{ij} = \frac{h_i \overline{\chi_j(g_i)}}{|G|}.$$

Proof. If B is the matrix whose ij entry is displayed in the statement, then

$$\begin{aligned}(AB)_{ij} &= \frac{1}{|G|} \sum_k \chi_i(g_k) h_k \overline{\chi_j(g_k)} \\ &= \frac{1}{|G|} \sum_g \chi_i(g) \overline{\chi_j(g)} \\ &= (\chi_i, \chi_j) \\ &= \delta_{ij},\end{aligned}$$

because $h_k \overline{\chi_j(g_k)} = \sum_{y \in C_k} \overline{\chi_j(y)}$. Therefore, $AB = I$. •

The next result is fundamental.

Theorem 8.135 (Orthogonality Relations). *Let G be a finite group of order n with conjugacy classes $C_1, \ldots, C_r$ of cardinalities $h_1, \ldots, h_r$, respectively, and choose elements*

$g_i \in C_i$. Let the irreducible characters of G be $\chi_1, \ldots, \chi_r$, and let χ_i have degree n_i. Then the following relations hold:

(i)

$$\sum_{k=1}^{r} h_k \chi_i(g_k)\overline{\chi_j(g_k)} = \begin{cases} 0 & \textit{if } i \neq j; \\ |G| & \textit{if } i = j. \end{cases}$$

(ii)

$$\sum_{i=1}^{r} \chi_i(g_k)\overline{\chi_i(g_\ell)} = \begin{cases} 0 & \textit{if } k \neq \ell; \\ |G|/h_k & \textit{if } k = \ell. \end{cases}$$

Proof. (i) This is just a restatement of Theorem 8.129.

(ii) If A is the character table of G and $B = [h_i\overline{\chi_j(g_i)}/|G|]$, we proved, in Lemma 8.134, that $AB = I$. It follows that $BA = I$; that is, $(BA)_{k\ell} = \delta_{k\ell}$. Therefore,

$$\frac{1}{|G|}\sum_i h_k\overline{\chi_i(g_k)}\chi_i(g_\ell) = \delta_{k\ell},$$

and this is the second orthogonality relation. •

In terms of the character table, the second orthogonality relation says that the usual (unweighted, but with complex conjugation) inner product of distinct columns is 0 while, for every k, the usual inner product of column k with itself is $|G|/h_k$.

The orthogonality relations yield the following special cases.

Corollary 8.136.

(i) $|G| = \sum_{i=1}^{r} n_i^2$

(ii) $\sum_{i=1}^{r} n_i\chi_i(g_k) = 0 \quad \textit{if } k > 1$

(iii) $\sum_{k=1}^{r} h_k\chi_i(g_k) = 0 \quad \textit{if } i > 1$

(iv) $\sum_{k=1}^{r} h_k|\chi_i(g_k)|^2 = |G|$

Proof. (i) This equation records the inner product of column 1 with itself: It is Theorem 8.135(ii) when $k = \ell = 1$.

(ii) This is the special case of Theorem 8.135(ii) with $\ell = 1$, for $\chi_i(1) = n_i$.

(iii) This is the special case of Theorem 8.135(i) in which $j = 1$.

(iv) This is the special case of Theorem 8.135(i) in which $j = i$. •

We can now give another proof of Burnside's lemma, Theorem 2.113, which counts the number of orbits of a G-set.

Theorem 8.137 (Burnside's Lemma). *Let G be a finite group and let X be a finite G-set. If N is the number of orbits of X, then*

$$N = \frac{1}{|G|}\sum_{g \in G} \mathrm{Fix}(g),$$

where $\mathrm{Fix}(g)$ *is the number of* $x \in X$ *with* $gx = x$.

Proof. Let V be the complex vector space having X as a basis. As in Example 8.117, the G-set X gives a representation $\sigma\colon G \to \mathrm{GL}(V)$ by $\sigma(g)(x) = gx$ for all $g \in G$ and $x \in X$; moreover, if χ_σ is the character afforded by σ, then Example 8.123(v) shows that $\chi_\sigma(g) = \mathrm{Fix}(g)$.

Let $O_1, \ldots, O_N$ be the orbits of X. We begin by showing that $N = \dim(V^G)$, where V^G is the space of *fixed points*:

$$V^G = \{v \in V : gv = v \text{ for all } g \in G\}.$$

For each i, define s_i to be the sum of all the x in O_i; it suffices to prove that these elements form a basis of V^G. It is plain that $s_1, \ldots, s_N$ is a linearly independent list in V^G, and it remains to prove that they span V^G. If $u \in V^G$, then $u = \sum_{x \in X} c_x x$, so that $gu = \sum_{x \in X} c_x (gx)$. Since $gu = u$, however, $c_x = c_{gx}$. Thus, given $x \in X$ with $x \in O_j$, each coefficient of gx, where $g \in G$, is equal to c_x; that is, all the x lying in the orbit O_j have the same coefficient, say, c_j, and so $u = \sum_j c_j s_j$. Therefore,

$$N = \dim(V^G).$$

Now define a linear transformation $T\colon V \to V$ by

$$T = \frac{1}{|G|}\sum_{g \in G} \sigma(g).$$

It is routine to check that T is a $\mathbb{C}G$-map, that $T|(V^G) =$ identity, and that $\mathrm{im}\, T = V^G$. Since $\mathbb{C}G$ is semisimple, $V = V^G \oplus W$ for some submodule W. We claim that $T|W = 0$. If $w \in W$, then $\sigma(g)(w) \in W$ for all $g \in G$, because W is a submodule, and so $T(w) \in W$. On the other hand, $T(w) \in \mathrm{im}\, T = V^G$, and so $T(w) \in V^G \cap W = \{0\}$, as claimed.

If $w_1, \ldots, w_\ell$ is a basis of W, then $s_1, \ldots, s_N, w_1, \ldots, w_\ell$ is a basis of $V = V^G \oplus W$. Note that T fixes each s_i and annihilates each w_j. Since trace preserves sums,

$$\begin{aligned}\mathrm{tr}(T) &= \frac{1}{|G|}\sum_{g \in G} \mathrm{tr}(\sigma(g)) \\ &= \frac{1}{|G|}\sum_{g \in G} \chi_\sigma(g) \\ &= \frac{1}{|G|}\sum_{g \in G} \mathrm{Fix}(g).\end{aligned}$$

It follows that

$$\operatorname{tr}(T) = \dim(V^G),$$

for the matrix of T with respect to the chosen basis is the direct sum of an identity block and a zero block, and so $\operatorname{tr}(T)$ is the size of the identity block, namely, $\dim(V^G) = N$. Therefore,

$$N = \frac{1}{|G|}\sum_{g\in G}\operatorname{Fix}(g). \quad \bullet$$

Character tables can be used to detect normal subgroups.

Definition. If χ_τ is the character afforded by a representation $\tau\colon G \to \mathrm{GL}(V)$, then

$$\ker\chi_\tau = \ker\tau.$$

Proposition 8.138. *Let $\theta = \chi_\tau$ be the character of a finite group G afforded by a representation $\tau\colon G \to \mathrm{GL}(V)$.*

(i) *For each $g \in G$, we have*

$$|\theta(g)| \le \theta(1).$$

(ii)

$$\ker\theta = \{g \in G : \theta(g) = \theta(1)\}.$$

(iii) *If $\theta = \sum_j m_j\chi_j$, where m_j are positive integers, then*

$$\ker\theta = \bigcap_j \ker\chi_j.$$

(iv) *If N is a normal subgroup of G, then there are irreducible characters $\chi_{i_1}, \ldots, \chi_{i_s}$ with $N = \bigcap_{j=1}^{s}\ker\chi_{i_j}$.*

Proof. (i) By Lagrange's theorem, $g^{|G|} = 1$ for every $g \in G$; it follows that the eigenvalues $\varepsilon_1, \ldots, \varepsilon_d$ of $\tau(g)$, where $d = \theta(1)$, are $|G|$th roots of unity, and so $|\varepsilon_j| = 1$ for all j. By the triangle inequality in $\mathbb{C}$,

$$\big|\theta(g)\big| = \Big|\sum_{j=1}^{d}\varepsilon_j\Big| \le d = \theta(1).$$

(ii) If $g \in \ker\theta = \ker\tau$, then $\tau(g) = I$, the identity matrix, and $|\theta(g)| = \operatorname{tr}(I) = \theta(1)$. Conversely, suppose that $\theta(g) = \theta(1) = d$; that is, $|\sum_{j=1}^{d}\varepsilon_j| = d$. By Proposition 1.42, all the eigenvalues ε_j are equal, say, $\varepsilon_j = \omega$ for all j. Therefore, $\tau(g) = \omega I$, by Corollary 8.120(iii), and so

$$\theta(g) = \theta(1)\omega.$$

But $\theta(g) = \theta(1)$, by hypothesis, and so $\omega = 1$; that is, $\tau(g) = I$ and $g \in \ker\tau$.

(iii) For all $g \in G$, we have

$$\theta(g) = \sum_j m_j \chi_j(g);$$

in particular,

$$\theta(1) = \sum_j m_j \chi_j(1).$$

If $g \in \ker\theta$, then $\theta(g) = \theta(1)$. Suppose that $\chi_{j'}(g) \neq \chi_{j'}(1)$ for some j'. Since $\chi_{j'}(g)$ is a sum of roots of unity, Proposition 1.42 applies to force $|\chi_{j'}(g)| < \chi_{j'}(1)$, and so $\theta(g) = \sum_j m_j \chi_j(g) \neq \theta(1)$. Therefore, $g \in \bigcap_j \ker\chi_j$. For the reverse inclusion, if $g \in \ker\chi_j$, then $\chi_j(g) = \chi_j(1)$, and so

$$\theta(g) = \sum_j m_j \chi_j(g) = \sum_j m_j \chi_j(1) = \theta(1);$$

hence, $g \in \ker\theta$.

(iv) It suffices to find a representation of G whose kernel is N. By part (iii) and Example 8.125(ii), the regular representation ρ of G/N is faithful (i.e., is an injection), and so its kernel is $\{1\}$. If $\pi : G \to G/N$ is the natural map, then $\rho\pi$ is a representation of G having kernel N. If θ is the character afforded by $\rho\pi$, then $\theta = \sum_j m_j \chi_j$, where the m_j are positive integers, by Lemma 8.126, and so part (iii) applies. •

Example 8.139.
We will construct the character table of S_4 in Example 8.148. We can see there that $\ker\chi_2 = A_4$ and $\ker\chi_3 = \mathbf{V}$ are the only two normal subgroups of S_4 (other than $\{1\}$ and S_4). Moreover, we can see that $\mathbf{V} \leq A_4$.

In Example 8.140, we can see that $\ker\chi_2 = \{1\} \cup z^G \cup y^G$ (where z^G denotes the conjugacy class of z in G) and $\ker\chi_3 = \{1\} \cup z^G \cup x^G$. Another normal subgroup occurs as $\ker\chi_2 \cap \ker\chi_3 = \{1\} \cup z^G$. ◂

A normal subgroup described by characters is given as a union of conjugacy classes; this viewpoint can give another proof of the simplicity of A_5. In Exercise 2.89(iv) on page 113, we saw that A_5 has five conjugacy classes, of sizes 1, 12, 12, 15, and 20. Since every subgroup contains the identity element, the order of a normal subgroup of A_5 is the sum of some of these numbers, including 1. But it is easy to see that 1 and 60 are the only such sums that are divisors of 60, and so the only normal subgroups are $\{1\}$ and A_5 itself.

There is a way to "lift" a representation of a quotient group to a representation of the group.

Definition. Let $H \lhd G$ and let $\sigma : G/H \to \mathrm{GL}(V)$ be a representation. If $\pi : G \to G/H$ is the natural map, then the representation $\sigma\pi : G \to \mathrm{GL}(V)$ is called a ***lifting*** of σ.

Scalar multiplication of G on a $\mathbb{C}G$-module V is given, for $v \in V$, by

$$gv = (gH)v.$$

Thus, every $\mathbb{C}G$-submodule of V is also a $\mathbb{C}(G/H)$-submodule; hence, if V is a simple $\mathbb{C}(G/H)$-module, then it is also a simple $\mathbb{C}G$-module. It follows that if $\sigma \colon G/H \to \mathrm{GL}(V)$ is an irreducible representation of G/H, then its lifting $\sigma\pi$ is also an irreducible representation of G.

Example 8.140.
We know that D_8 and $\mathbf{Q}$ are nonisomorphic nonabelian groups of order 8; we now show that they have the same character tables.

If G is a nonabelian group of order 8, then its center has order 2, say, $Z(G) = \langle z \rangle$. Now $G/Z(G)$ is not cyclic, by Exercise 2.69 on page 95, and so $G/Z(G) \cong \mathbf{V}$. Therefore, if $\sigma \colon \mathbf{V} \to \mathbb{C}$ is an irreducible representation of $\mathbf{V}$, then its lifting $\sigma\pi$ is an irreducible representation of G. This gives 4 (necessarily irreducible) linear characters of G, each of which takes value 1 on z. As G is not abelian, there must be an irreducible character χ_5 of degree $n_5 > 1$ (if all $n_i = 1$, then $\mathbb{C}G$ is commutative and G is abelian). Since $\sum_i n_i^2 = 8$, we see that $n_5 = 2$. Thus, there are five irreducible representations and, hence, five conjugacy classes; choose representatives g_i to be $1, z, x, y, w$. Table 8.4 is the character table.

g_i	1	z	x	y	w
h_i	1	1	2	2	2
χ_1	1	1	1	1	1
χ_2	1	1	-1	1	-1
χ_3	1	1	1	-1	-1
χ_4	1	1	-1	-1	1
χ_5	2	-2	0	0	0

Table 8.4. Character Table of D_8 and of $\mathbf{Q}$

The values for χ_5 are computed from the orthogonality relations of the columns. For example, if the last row of the character table is

$$2 \quad a \quad b \quad c \quad d,$$

then the inner product of columns 1 and 2 gives the equation $4 + 2a = 0$, so that $a = -2$. The reader may verify that $0 = b = c = d$. ◂

The orthogonality relations help to complete a character table but, obviously, it would also be useful to have a supply of characters. One important class of characters consists of those afforded by *induced representations*; that is, representations of a group G determined by representations of a subgroup H of G.

The original construction of induced representations, due to F. G. Frobenius, is rather complicated. Tensor products make this construction more natural. The ring $\mathbb{C}G$ is a $(\mathbb{C}G, \mathbb{C}H)$-bimodule (for $\mathbb{C}H$ is a subring of $\mathbb{C}G$), so that if V is a left $\mathbb{C}H$-module, then the tensor product $\mathbb{C}G \otimes_{\mathbb{C}H} V$ is defined; Lemma 8.80 says that this tensor product is, in fact, a left $\mathbb{C}G$-module.

Definition. Let H be a subgroup of a group G. If V is a left $\mathbb{C}H$-module, then the ***induced module*** is the left $\mathbb{C}G$-module

$$V\!\upharpoonleft^G = \mathbb{C}G \otimes_{\mathbb{C}H} V.$$

The corresponding representation $\rho\!\upharpoonleft^G : G \to V^G$ is called the ***induced representation***. The character of G afforded by $\rho\!\upharpoonleft^G$ is called the ***induced character***, and it is denoted by $\chi_\rho\!\upharpoonleft^G$.

Let us recognize at the outset that the character of an induced representation need not restrict to the original representation of the subgroup. For example, we have seen that there is an irreducible character χ of $A_3 \cong \mathbb{I}_3$ having complex values, whereas every irreducible character of S_3 has (real) integer values. A related observation is that two elements may be conjugate in a group but not conjugate in a subgroup (for example, nontrivial elements in A_3 are conjugate in S_3, for they have the same cycle structure, but they are not conjugate in the abelian group A_3).

The next lemma will help us compute the character afforded by an induced representation.

Lemma 8.141.

(i) *If $H \le G$, then $\mathbb{C}G$ is a free right $\mathbb{C}H$-module on $[G : H]$ generators.*

(ii) *If a left $\mathbb{C}H$-module V has a (vector space) basis $e_1, \ldots, e_m$, then a (vector space) basis of the induced module $V\!\upharpoonleft^G = \mathbb{C}G \otimes_{\mathbb{C}H} V$ is the family of all $t_i \otimes e_j$, where $t_1, \ldots, t_n$ is a transversal of H in G.*

Proof. (i) Since $t_1, \ldots, t_n$ is a transversal of H in G (of course, $n = [G : H]$), we see that G is the disjoint union

$$G = \bigcup_i t_i H;$$

thus, for every $g \in G$, there is a unique i and a unique $h \in H$ with $g = t_i h$. We claim that $t_1, \ldots, t_n$ is a basis of $\mathbb{C}G$ viewed as a right $\mathbb{C}H$-module.

If $u \in \mathbb{C}G$, then $u = \sum_g a_g g$, where $a_g \in \mathbb{C}$. Rewrite each term

$$a_g g = a_g t_i h = t_i a_g h$$

(scalars in $\mathbb{C}$ commute with everything), collect terms involving the same t_i, and obtain $u = \sum_i t_i \eta_i$, where $\eta_i \in \mathbb{C}H$.

To prove uniqueness of this expression, suppose that $0 = \sum_i t_i \eta_i$, where $\eta_i \in \mathbb{C}H$. Now $\eta_i = \sum_{h \in H} a_{ih} h$, where $a_{ih} \in \mathbb{C}$. Substituting,

$$0 = \sum_{i,h} a_{ih} t_i h.$$

But $t_i h = t_j h'$ if and only if $i = j$ and $h = h'$, so that $0 = \sum_{i,h} a_{ih} t_i h = \sum_{g \in G} a_{ih} g$, where $g = t_i h$. Since the elements of G are a basis of $\mathbb{C}G$ (viewed as a vector space over $\mathbb{C}$), we have $a_{ih} = 0$ for all i, h, and so $\eta_i = 0$ for all i.

(ii) By Theorem 8.87,

$$\mathbb{C}G \otimes_{\mathbb{C}H} V \cong \sum_i t_i \mathbb{C}H \otimes_{\mathbb{C}H} V.$$

It follows that every $u \in \mathbb{C}G \otimes_{\mathbb{C}H} V$ has a unique expression as a $\mathbb{C}$-linear combination of $t_i \otimes e_j$, and so these elements comprise a basis. •

We introduce the following notation. If $H \leq G$ and $\chi \colon H \to \mathbb{C}$ is a function, then $\dot{\chi} \colon G \to \mathbb{C}$ is given by

$$\dot{\chi}(g) = \begin{cases} 0 & \text{if } g \notin H \\ \chi(g) & \text{if } g \in H. \end{cases}$$

Theorem 8.142. *If χ_σ is the character afforded by a representation $\sigma \colon H \to \mathrm{GL}(V)$ of a subgroup H of a group G, then the induced character $\chi_\sigma\!\uparrow^G$ is given by*

$$\chi_\sigma\!\uparrow^G(g) = \frac{1}{|H|} \sum_{a \in G} \dot{\chi}_\sigma(a^{-1} g a).$$

Proof. Let $t_1, \ldots, t_n$ be a transversal of H in G, so that G is the disjoint union $G = \bigcup_i t_i H$, and let $e_1, \ldots, e_m$ be a (vector space) basis of V. By Lemma 8.141(ii), a basis for the vector space $V^G = \mathbb{C}G \otimes_{\mathbb{C}H} V$ consists of all $t_i \otimes e_j$. If $g \in G$, we compute the matrix of left multiplication by g relative to this basis. Note that

$$g t_i = t_{k(i)} h_i,$$

where $h_i \in H$, and so

$$\begin{aligned} g(t_i \otimes e_j) &= (g t_i) \otimes e_j \\ &= t_{k(i)} h_i \otimes e_j \\ &= t_{k(i)} \otimes \sigma(h_i) e_j \end{aligned}$$

(the last equation holds because we can slide any element of H across the tensor sign). Now $g(t_i \otimes e_j)$ is written as a $\mathbb{C}$-linear combination of *all* the basis elements of $V\!\uparrow^G$, for the coefficients $t_p \otimes e_j$ for $p \neq k(i)$ are all 0. Hence, $\sigma\!\uparrow^G(g)$ gives the $nm \times nm$ matrix whose m columns labeled by $t_i \otimes e_j$, for fixed i, are all zero except for an $m \times m$ block equal to

$$[a_{pq}(h_i)] = [a_{pq}(t_{k(i)}^{-1} g t_i)].$$

Thus, the big matrix is partitioned into $m \times m$ blocks, most of which are 0, and a nonzero block is on the diagonal of the big matrix if and only if $k(i) = i$; that is,

$$t_{k(i)}^{-1} g t_i = t_i^{-1} g t_i = h_i \in H.$$

The induced character is the trace of the big matrix, which is the sum of the traces of these blocks on the diagonal. Therefore,

$$\begin{aligned} \chi_\sigma\!\uparrow^G(g) &= \sum_{t_i^{-1} g t_i \in H} \operatorname{tr}([a_{pq}(t_i^{-1} g t_i)]) \\ &= \sum_i \dot{\chi}_\sigma(t_i^{-1} g t_i) \end{aligned}$$

(remember that $\dot{\chi}_\sigma$ is 0 outside of H). We now rewrite the summands (to get a formula that does not depend on the choice of the transversal): If $t_i^{-1} g t_i \in H$, then $(t_i h)^{-1} g (t_i h) = h^{-1}(t_i^{-1} g t_i) h$ in H, so that, for fixed i,

$$\sum_{h \in H} \dot{\chi}_\sigma\big((t_i h)^{-1} g (t_i h)\big) = |H| \dot{\chi}_\sigma(t_i^{-1} g t_i),$$

because χ_σ is a class function on H. Therefore,

$$\begin{aligned} \chi_\sigma\!\uparrow^G(g) &= \sum_i \dot{\chi}_\sigma(t_i^{-1} g t_i) \\ &= \frac{1}{|H|} \sum_{i,h} \dot{\chi}_\sigma\big((t_i h)^{-1} g (t_i h)\big) \\ &= \frac{1}{|H|} \sum_{a \in G} \dot{\chi}_\sigma(a^{-1} g a). \quad \bullet \end{aligned}$$

Remark. We have been considering induced characters, but it is easy to generalize the discussion to ***induced class functions***. If $H \le G$, then a class function θ on H has a unique expression as a $\mathbb{C}$-linear combination of irreducible characters of H, say, $\theta = \sum c_i \chi_i$, and so we can define

$$\theta\!\uparrow^G = \sum c_i \chi_i\!\uparrow^G.$$

It is plain that $\theta\!\uparrow^G$ is a class function on G, and that the formula in Theorem 8.142 extends to induced class functions. ◀

If, for $h \in H$, the matrix of $\sigma(h)$ (with respect to the basis $e_1, \ldots, e_m$ of V) is $B(h)$, then define $m \times m$ matrices $\dot{B}(g)$, for all $g \in G$, by

$$\dot{B}(g) = \begin{cases} 0 & \text{if } g \notin H; \\ B(g) & \text{if } g \in H. \end{cases}$$

The proof of Theorem 8.142 allows us to picture the matrix of the induced representation in block form

$$\sigma\!\uparrow^G(g) = \begin{bmatrix} \dot{B}(t_1^{-1}gt_1) & \dot{B}(t_1^{-1}gt_2) & \cdots & \dot{B}(t_1^{-1}gt_n) \\ \dot{B}(t_2^{-1}gt_1) & \dot{B}(t_2^{-1}gt_2) & \cdots & \dot{B}(t_2^{-1}gt_n) \\ \vdots & \vdots & \vdots & \vdots \\ \dot{B}(t_n^{-1}gt_1) & \dot{B}(t_n^{-1}gt_2) & \cdots & \dot{B}(t_n^{-1}gt_n) \end{bmatrix}.$$

Corollary 8.143. *Let H be a subgroup of a finite group G and let χ be a character on H.*

(i) $\chi\!\uparrow^G(1) = [G : H]\chi(1)$.

(ii) *If $H \lhd G$, then $\chi\!\uparrow^G(g) = 0$ for all $g \notin H$.*

Proof. (i) For all $a \in G$, we have $a^{-1}1a = 1$, so that there are $|G|$ terms in the sum $\chi\!\uparrow^G(1) = \frac{1}{|H|}\sum_{a\in G}\dot{\chi}(a^{-1}ga)$ that are equal to $\chi(1)$; hence,

$$\chi\!\uparrow^G(1) = \frac{|G|}{|H|}\chi(1) = [G : H]\chi(1).$$

(ii) If $H \lhd G$, then $g \notin H$ implies $a^{-1}ga \notin H$ for all $a \in G$. Therefore, $\dot{\chi}(a^{-1}ga) = 0$ for all $a \in G$, and so $\chi\!\uparrow^G(g) = 0$. •

Example 8.144.
Let $H \le G$ be a subgroup of index n, let $X = \{t_1H, \ldots, t_nH\}$ be the family of left cosets of H, and let $\varphi\colon G \to S_X$ be the (permutation) representation of G on the cosets of H. As in Example 8.123(v), we may regard $\varphi\colon G \to \mathrm{GL}(V)$, where V is the vector space over $\mathbb{C}$ having basis X; that is, φ is a representation in the sense of this section.

We claim that if χ_φ is the character afforded by φ, then $\chi_\varphi = \epsilon\!\uparrow^G$, where ϵ is the trivial character on H. On the one hand, Example 8.123(v) shows that

$$\chi_\varphi(g) = \mathrm{Fix}(\varphi(g))$$

for every $g \in G$. On the other hand, suppose $\varphi(g)$ is the permutation (in two-rowed notation)

$$\varphi(g) = \begin{pmatrix} t_1H & \ldots & t_nH \\ gt_1H & \ldots & gt_nH \end{pmatrix}.$$

Now $gt_iH = t_iH$ if and only if $t_i^{-1}gt_i \in H$. Thus, $\dot{\epsilon}(t_i^{-1}gt_i) \neq 0$ if and only if $gt_iH = t_iH$, and so

$$\epsilon\!\uparrow^G(g) = \mathrm{Fix}(\varphi(g)). \quad ◂$$

Even though a character λ of a subgroup H is irreducible, its induced character need not be irreducible. For example, let $G = S_3$ and H be the cyclic subgroup generated by $(1\ 2)$. The linear representation $\sigma = \mathrm{sgn}\colon H \to \mathbb{C}$ is irreducible, and it affords the character χ_σ with

$$\chi_\sigma(1) = 1 \quad \text{and} \quad \chi_\sigma((1\ 2)) = -1.$$

Using the formula for the induced character, we find that

$$\chi_\sigma{\uparrow}^{S_3}(1) = 3, \quad \chi_\sigma{\uparrow}^{S_3}((1\ 2)) = -1, \quad \text{and} \quad \chi_\sigma{\uparrow}^{S_3}((1\ 2\ 3)) = 0.$$

Corollary 8.130 shows that $\chi_\sigma{\uparrow}^{S_3}$ is not irreducible, for $(\chi_\sigma{\uparrow}^{S_3}, \chi_\sigma{\uparrow}^{S_3}) = 2$. It is easy to see that $\chi_\sigma{\uparrow}^{S_3} = \chi_2 + \chi_3$, the latter being the nontrivial irreducible characters of S_3.

We must mention a result of R. Brauer. Call a subgroup E of a finite group G ***elementary*** if $E = Z \times P$, where Z is cyclic and P is a p-group for some prime p.

Theorem (Brauer). *Every complex character θ on a finite group G has the form*

$$\theta = \sum_i m_i \mu_i{\uparrow}^G,$$

where $m_i \in \mathbb{Z}$ and the μ_i are linear characters on elementary subgroups of G.

Proof. See Curtis–Reiner, *Representation Theory of Finite Groups and Associative Algebras*, page 283. •

Definition. If H is a subgroup of a group G, then every representation $\sigma\colon G \to \mathrm{GL}(V)$ gives, by restriction, a representation $\sigma|H\colon H \to \mathrm{GL}(V)$. (In terms of modules, every left $\mathbb{C}G$-module V can be viewed as a left $\mathbb{C}H$-module.) We call $\sigma|H$ the ***restriction*** of σ, and we denote it by $\sigma{\downarrow}_H$. The character of H afforded by $\sigma{\downarrow}_H$ is denoted by $\chi_\sigma{\downarrow}_H$.

The next result displays an interesting relation between characters on a group and characters on a subgroup. (Formally, it resembles the adjoint isomorphism.)

Theorem 8.145 (Frobenius Reciprocity). *Let H be a subgroup of a group G, let χ be a class function on G, and let θ be a class function on H. Then*

$$(\theta{\uparrow}^G, \chi)_G = (\theta, \chi{\downarrow}_H)_H,$$

where $(\ ,\ )_G$ denotes the inner product on $\mathrm{cf}(G)$ *and $(\ ,\ )_H$ denotes the inner product on* $\mathrm{cf}(H)$.

Proof.

$$\begin{aligned}(\theta\!\uparrow^G, \chi)_G &= \frac{1}{|G|}\sum_{g\in G}\theta\!\uparrow^G(g)\overline{\chi(g)}\\ &= \frac{1}{|G|}\sum_{g\in G}\frac{1}{|H|}\sum_{a\in G}\dot{\theta}(a^{-1}ga)\overline{\chi(g)}\\ &= \frac{1}{|G|}\frac{1}{|H|}\sum_{a,g\in G}\dot{\theta}(a^{-1}ga)\overline{\chi(a^{-1}ga)},\end{aligned}$$

the last equation occurring because χ is a class function. For fixed $a \in G$, as g ranges over G, then so does $a^{-1}ga$. Therefore, writing $x = a^{-1}ga$, the equations continue

$$\begin{aligned}&= \frac{1}{|G|}\frac{1}{|H|}\sum_{a,x\in G}\dot{\theta}(x)\overline{\chi(x)}\\ &= \frac{1}{|G|}\frac{1}{|H|}\sum_{a\in G}\Big(\sum_{x\in G}\dot{\theta}(x)\overline{\chi(x)}\Big)\\ &= \frac{1}{|G|}\frac{1}{|H|}|G|\sum_{x\in G}\dot{\theta}(x)\overline{\chi(x)}\\ &= \frac{1}{|H|}\sum_{x\in G}\dot{\theta}(x)\overline{\chi(x)}\\ &= (\theta, \chi\!\downarrow_H)_H,\end{aligned}$$

the next to last equation holding because $\dot{\theta}(x)$ vanishes off the subgroup H. •

The following elementary remark facilitates the computation of induced class functions.

Lemma 8.146. *Let H be a subgroup of a finite group G, and let χ be a class function on H. Then*

$$\chi\!\uparrow^G(g) = \frac{1}{|H|}\sum_i |C_G(g_i)|\dot{\chi}(g_i^{-1}gg_i).$$

Proof. Let $|C_G(g_i)| = m_i$. If $a_0^{-1}g_ia_0 = g$, we claim that there are exactly m_i elements $a \in G$ with $a^{-1}g_ia = g$. There are at least m_i elements in G conjugating g_i to g; namely, all aa_0 for $a \in C_G(g_i)$. There are at most m_i elements, for if $b^{-1}g_ib = g$, then $b^{-1}g_ib = a_0^{-1}g_ia_0$, and so $a_0b \in C_G(g_i)$. The result now follows by collecting terms involving g_is in the formula for $\chi\!\uparrow^G(g)$. •

Example 8.147.
Table 8.5 on page 630 is the character table of A_4, where $\omega = e^{2\pi i/3}$ is a primitive cube root of unity.

g_i	(1)	(1 2 3)	(1 3 2)	(1 2)(3 4)
h_i	1	4	4	3
χ_1	1	1	1	1
χ_2	1	ω	ω^2	1
χ_3	1	ω^2	ω	1
χ_4	3	0	0	-1

Table 8.5. Character Table of A_4

The group A_4 consists of the identity, eight 3-cycles, and three products of disjoint transpositions. In S_4, all the 3-cycles are conjugate; if $g = (1\ 2\ 3)$, then $[S_4 : C_{S_4}(g)] = 8$. It follows that $|C_{S_4}(g)| = 3$, and so $C_{S_4}(g) = \langle g \rangle$. Therefore, in A_4, the number of conjugates of g is $[A_4 : C_{A_4}(g)] = 4$ [we know that $C_{A_4}(g) = A_4 \cap C_{S_4}(g) = \langle g \rangle$]. The reader may show that g and g^{-1} are not conjugate, and so we have verified the first two rows of the character table.

The rows for χ_2 and χ_3 are liftings of linear characters of $A_4/\mathbf{V} \cong \mathbb{I}_3$. Note that if $h = (1\ 2)(3\ 4)$, then $\chi_2(h) = \chi_2(1) = 1$, because $\mathbf{V}$ is the kernel of the lifted representation; similarly, $\chi_3(h) = 1$. Now $\chi_4(1) = 3$, because $3 + (n_4)^2 = 12$. The bottom row arises from orthogonality of the columns. (We can check, using Corollary 8.130, that the character of degree 3 is irreducible.) ◂

Example 8.148.
Table 8.6 is the character table of S_4.

g_i	(1)	(1 2)	(1 2 3)	(1 2 3 4)	(1 2)(3 4)
h_i	1	6	8	6	3
χ_1	1	1	1	1	1
χ_2	1	-1	1	-1	1
χ_3	2	0	-1	0	2
χ_4	3	1	0	-1	-1
χ_5	3	-1	0	1	-1

Table 8.6. Character Table of S_4

We know, for all n, that two permutations in S_n are conjugate if and only if they have the same cycle structure; the sizes of the conjugacy classes in S_4 were computed in Example 2.5(i).

The rows for χ_2 and χ_3 are liftings of irreducible characters of $S_4/\mathbf{V} \cong S_3$. The entries in the fourth column of these rows arise from $(1\ 2)\mathbf{V} = (1\ 2\ 3\ 4)\mathbf{V}$; the entries in the last column of these rows arise from $\mathbf{V}$ being the kernel (in either case), so that $\chi_j((1\ 2)(3\ 4)) = \chi_j(1)$ for $j = 2, 3$.

We complete the first column using $24 = 1 + 1 + 4 + n_4^2 + n_5^2$; thus, $n_4 = 3 = n_5$. Let us see whether χ_4 is an induced character; if it is, then Corollary 8.143(i) shows that

it arises from a linear character of a subgroup H of index 3. Such a subgroup has order 8, and so it is a Sylow 2-subgroup; that is, $H \cong D_8$. Let us choose one such subgroup: Let

$$H = \langle \mathbf{V}, (1\ 3) \rangle = \mathbf{V} \cup \{(1\ 3), (2\ 4), (1\ 2\ 3\ 4), (1\ 4\ 3\ 2)\}.$$

The conjugacy classes are

$$\begin{aligned} C_1 &= \{1\}; \\ C_2 &= \{(1\ 3)(2\ 4)\}; \\ C_3 &= \{(1\ 2)(3\ 4), (1\ 4)(2\ 3)\}; \\ C_4 &= \{(1\ 3), (2\ 4)\}; \\ C_5 &= \{(1\ 2\ 3\ 4), (1\ 4\ 3\ 2)\}. \end{aligned}$$

Let θ be the character on H defined by

C_1	C_2	C_3	C_4	C_5
1	1	-1	1	-1.

Define $\chi_4 = \theta\!\uparrow^{S_4}$. Using the formula for induced characters, assisted by Lemma 8.146, we obtain the fourth row of the character table. However, before going on to row 5, we observe that Corollary 8.130 shows that χ_4 is irreducible, for $(\chi_4, \chi_4) = 1$. Finally, the orthogonality relations allows us to compute row 5. ◀

At this point in the story, we must introduce algebraic integers. Since G is a finite group, Lagrange's theorem gives $g^{|G|} = 1$ for all $g \in G$. It follows that if $\sigma : G \to \mathrm{GL}(V)$ is a representation, then $\sigma(g)^{|G|} = I$ for all g; hence, all the eigenvalues of $\sigma(g)$ are $|G|$th roots of unity, and so all the eigenvalues are algebraic integers. By Proposition 7.24, the trace of $\sigma(g)$, being the sum of the eigenvalues, is also an algebraic integer.

We can now prove the following interesting result.

Theorem 8.149. *The degrees n_i of the irreducible characters of a finite group G are divisors of $|G|$.*

Proof. By Corollary 3.44, the rational number $\alpha = |G|/n_i$ is an integer if it is also an algebraic integer. Now Corollary 8.10(ii) says that α is an algebraic integer if there is a faithful $\mathbb{Z}[\alpha]$-module M that is a finitely generated abelian group, where $\mathbb{Z}[\alpha]$ is the smallest subring of $\mathbb{C}$ containing α.

By Proposition 8.128, we have

$$\begin{aligned} e_i &= \sum_{g \in G} \frac{n_i}{|G|} \chi_i(g^{-1}) g \\ &= \sum_{g \in G} \frac{1}{\alpha} \chi_i(g^{-1}) g. \end{aligned}$$

Hence, $\alpha e_i = \sum_{g \in G} \chi_i(g^{-1})g$. But e_i is an idempotent: $e_i^2 = e_i$, and so

$$\alpha e_i = \sum_{g \in G} \chi_i(g^{-1}) g e_i.$$

Define M to be the abelian subgroup of $\mathbb{C}G$ generated by all elements of the form $\zeta g e_i$, where ζ is a $|G|$th root of unity and $g \in G$; of course, M is a finitely generated abelian group.

To see that M is a $\mathbb{Z}[\alpha]$-module, it suffices to show that $\alpha M \subseteq M$. But

$$\begin{aligned} \alpha \zeta g e_i &= \zeta g \alpha e_i \\ &= \zeta g \sum_{h \in G} \chi_i(h^{-1}) h e_i \\ &= \sum_{h \in G} \chi_i(h^{-1}) \zeta g h e_i. \end{aligned}$$

This last element lies in M, however, because $\chi_i(h^{-1})$ is a sum of $|G|$th roots of unity.

Finally, if $\beta \in \mathbb{C}$ and $u \in \mathbb{C}G$, then $\beta u = 0$ if and only if $\beta = 0$ or $u = 0$. Since $\mathbb{Z}[\alpha] \subseteq \mathbb{C}$ and $M \subseteq \mathbb{C}G$, however, it follows that M is a faithful $\mathbb{Z}[\alpha]$-module, as desired. •

We will present two important applications of character theory in the next section; for other applications, as well as a more serious study of representations, the interested reader should look at the books of Curtis–Reiner, Feit, Huppert, and Isaacs. Representation theory was an essential ingredient of the proof of the classification of the finite simple groups in the 1980s: There are several infinite families and 26 ***sporadic*** groups belonging to no infinite family (see the chapter by R. Carter in the book edited by Kostrikin and Shafarevich, as well as Gorenstein–Lyons–Solomon, *The Classification of the Finite Simple Groups*). The *ATLAS*, by Conway et al, contains the character tables of every simple group of order under 10^{25} as well as the character tables of all the sporadic groups. The largest sporadic simple group is called the ***Monster***; it has order

$$2^{46} \cdot 3^{20} \cdot 5^9 \cdot 7^6 \cdot 11^2 \cdot 13^3 \cdot 17 \cdot 19 \cdot 23 \cdot 29 \cdot 31 \cdot 41 \cdot 47 \cdot 59 \cdot 71.$$

EXERCISES

8.55 Prove that if θ is a generalized character of a finite group G, then there are characters χ and ψ with $\theta = \chi - \psi$.

8.56 (i) Prove that if z is a complex root of unity, then $z^{-1} = \overline{z}$.

(ii) Prove that if G is a finite group and $\sigma : G \to \mathrm{GL}(V)$ is a representation, then

$$\chi_\sigma(g^{-1}) = \overline{\chi_\sigma(g)}$$

for all $g \in G$.

Hint. Use the fact that every eigenvalue of $\sigma(g)$ is a root of unity, as well as the fact that if A is a nonsingular matrix over a field k and if $u_1, \ldots, u_n$ are the eigenvalues of A (with multiplicities), then the eigenvalues of A^{-1} are $u_1^{-1}, \ldots, u_n^{-1}$; that is, $\overline{u_1}, \ldots, \overline{u_n}$.

8.57 If $\sigma \colon G \to \mathrm{GL}(n, \mathbb{C})$ is a representation, its ***contragredient*** $\sigma^* \colon G \to \mathrm{GL}(n, \mathbb{C})$ is the function given by

$$\sigma^*(g) = \sigma(g^{-1})^t,$$

where ^{t} denotes transpose.

(i) Prove that the contragredient of a representation σ is a representation that is irreducible when σ is irreducible.

(ii) Prove that the character χ_{σ^*} afforded by the contragredient σ^* is

$$\chi_{\sigma^*}(g) = \overline{\chi_\sigma(g)},$$

where $\overline{\chi_\sigma(g)}$ is the complex conjugate. Conclude that if χ is a character of G, then $\overline{\chi}$ is also a character.

8.58 Construct an irreducible representation of S_3 of degree 2.

8.59 (i) If $g \in G$, where G is a finite group, prove that g is conjugate to g^{-1} if and only if $\chi(g)$ is real for every character χ of G.

(ii) Prove that every character of S_n is real valued. (It is a theorem of F. G. Frobenius that every character of S_n is integer valued.)

8.60 (i) If G is a finite abelian group, define its ***character group*** G^* by

$$G^* = \mathrm{Hom}(G, \mathbb{C}^\times),$$

where $\mathbb{C}^\times$ is the multiplicative group of nonzero complex numbers. Prove that $G^* \cong G$.

Hint. Use the fundamental theorem of finite abelian groups.

(ii) Prove that $\mathrm{Hom}(G, \mathbb{C}^\times) \cong \mathrm{Hom}(G, \mathbb{Q}/\mathbb{Z})$ when G is a finite abelian group.

(iii) Prove that every irreducible character of a finite abelian group is linear.

8.61 Prove that the only linear character of a simple group is the trivial character. Conclude that if χ_i is not the trivial character, then $n_i = \chi_i(1) > 1$.

8.62 Let $\theta = \chi_\sigma$ be the character afforded by a representation σ of a finite group G.

(i) If $g \in G$, prove that $|\theta(g)| = \theta(1)$ if and only if $\sigma(g)$ is a scalar matrix.

Hint. Use Proposition 1.42 on page 23.

(ii) If θ is an irreducible character, prove that

$$Z(G/\ker\theta) = \{g \in G : |\theta(g)| = \theta(1)\}.$$

8.63 If G is a finite group, prove that the number of its (necessarily irreducible) linear representations is $[G : G']$.

8.64 Let G be a finite group.

(i) If $g \in G$, show that $|C_G(g)| = \sum_{i=1}^r |\chi_i(g)|^2$. Conclude that the character table of G gives $|C_G(g)|$.

(ii) Show how to use the character table of G to see whether G is abelian.

(iii) Show how to use the character table of G to find the lattice of normal subgroups of G and their orders.

(iv) If G is a finite group, show how to use its character table to find the commutator subgroup G'.

Hint. If $K \lhd G$, then the character table of G/K is a submatrix of the character table of G, and so we can find the abelian quotient of G having largest order.

(v) Show how to use the character table of a finite group G to determine whether G is solvable.

8.65 (i) Show how to use the character table of G to find $|Z(G)|$.

(ii) Show how to use the character table of a finite group G to determine whether G is nilpotent.

8.66 Recall that the group $\mathbf{Q}$ of quaternions has the presentation

$$\mathbf{Q} = (A, B | A^4 = 1, A^2 = B^2, BAB^{-1} = A^{-1}).$$

(i) Show that there is a representation $\sigma : \mathbf{Q} \to \mathrm{GL}(2, \mathbb{C})$ with

$$A \mapsto \begin{bmatrix} i & 0 \\ 0 & -i \end{bmatrix} \text{ and } B \mapsto \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}.$$

(ii) Prove that σ is an irreducible representation.

8.67 (i) If $\sigma : G \to \mathrm{GL}(V)$ and $\tau : G \to \mathrm{GL}(W)$ are representations, prove that

$$\sigma \otimes \tau : G \to \mathrm{GL}(V \otimes W),$$

defined by

$$(\sigma \otimes \tau)(g) = \sigma(g) \otimes \tau(g)$$

is a representation.

(ii) Prove that the character afforded by $\sigma \otimes \tau$ is the pointwise product:

$$\chi_\sigma \chi_\tau : g \mapsto \mathrm{tr}(\sigma(g))\, \mathrm{tr}(\tau(g)).$$

(iii) Prove that $\mathrm{cf}(G)$ is a commutative ring (usually called the ***Burnside ring*** of G).

8.6 THEOREMS OF BURNSIDE AND OF FROBENIUS

Character theory will be used in this section to prove two important results in group theory: Burnside's $p^m q^n$ theorem and a theorem of Frobenius. We begin with the following variation of Schur's lemma.

Proposition 8.150. *If $\sigma : G \to \mathrm{GL}(V)$ is an irreducible representation and if a linear transformation $\varphi : V \to V$ satisfies*

$$\varphi\sigma(g) = \sigma(g)\varphi$$

for all $g \in G$, then φ is a scalar transformation: there exists $\alpha \in \mathbb{C}$ with $\varphi = \alpha 1_V$.

Proof. The vector space V is a $\mathbb{C}G$-module with scalar multiplication $gv = \sigma(g)(v)$ for all $v \in V$, and any linear transformation θ satisfying the equation $\theta\sigma(g) = \sigma(g)\theta$ for all $g \in G$ is a $\mathbb{C}G$-map $V^\sigma \to V^\sigma$. Since σ is irreducible, the $\mathbb{C}G$-module V^σ is simple; by Schur's lemma [Theorem 8.52(ii)], we have $\text{End}(V^\sigma)$ a division ring, and so every nonzero element in it is nonsingular. Now $\varphi - \alpha 1_V \in \text{End}(V^\sigma)$ for every $\alpha \in \mathbb{C}$; in particular, this is so when α is an eigenvalue of φ (which lies in $\mathbb{C}$ because $\mathbb{C}$ is algebraically closed). The definition of eigenvalue says that $\varphi - \alpha 1_V$ is singular, and so it must be 0; that is, $\varphi = \alpha 1_V$, as desired. •

As in Proposition 8.119(ii), we may regard the irrreducible representation $\lambda_i : G \to \text{GL}(L_i)$, given by left multiplication on the minimal left ideal L_i, as a $\mathbb{C}$-algebra map $\widetilde{\lambda}_i : \mathbb{C}G \to \text{End}(L_i)$ (after all, $\text{im}\,\widetilde{\lambda}_i \subseteq \text{End}(L_i)$). Hence, the restriction to the center of $\mathbb{C}G$ is also an algebra map:

$$\widetilde{\lambda}_i : Z(\mathbb{C}G) \to \text{End}(L_i) \cong \text{Mat}_{n_i}(\mathbb{C}).$$

Thus, for each $z \in Z(\mathbb{C}G)$, we see that $\widetilde{\lambda}_i(z)$ is an $n_i \times n_i$ complex matrix. By Proposition 8.150, each $\widetilde{\lambda}_i(z)$ is a scalar matrix for every $z \in Z(\mathbb{C}G)$:

$$\widetilde{\lambda}_i(z) = \omega_i(z)I,$$

where $\omega_i(z) \in \mathbb{C}$. Moreover, the function $\omega_i : Z(\mathbb{C}G) \to \mathbb{C}$ is a $\mathbb{C}$-algebra map because $\widetilde{\lambda}_i$ is.

Recall, from Lemma 8.68, that a basis for $Z(\mathbb{C}G)$ consists of the *class sums*

$$z_i = \sum_{g \in C_i} g,$$

where the conjugacy classes of G are $C_1, \ldots, C_r$.

Proposition 8.151. *Let $z_1, \ldots, z_r$ be the class sums of a finite group G.*

(i) *For each i, j, we have*
$$\omega_i(z_j) = \frac{h_j \chi_i(g_j)}{n_i},$$
where $g_j \in C_j$.

(ii) *There are nonnegative integers $a_{ij\nu}$ with*
$$z_i z_j = \sum_\nu a_{ij\nu} z_\nu.$$

(iii) *The complex numbers $\omega_i(z_j)$ are algebraic integers.*

Proof. (i) Computing the trace of $\widetilde{\lambda}_i(z_j) = \omega_i(z_j)I$ gives

$$n_i\omega_i(z_j) = \chi_i(z_j) = \sum_{g \in C_j} \chi_i(g) = h_j\chi_i(g_j),$$

for χ_i is constant on the conjugacy class C_j. Therefore, $\omega_i(z_j) = h_j\chi_i(g_j)/n_i$.

(ii) Choose $g_\nu \in C_\nu$. The definition of multiplication in the group algebra shows that the coefficient of g_ν in $z_i z_j$ is

$$|\{(g_i, g_j) \in C_i \times C_j : g_i g_j = g_\nu\}|,$$

the cardinality of a finite set, and hence it is a nonnegative integer. As all the coefficients of z_ν are equal [for we are in $Z(\mathbb{C}G)$], it follows that this number is $a_{ij\nu}$.

(iii) Let M be the (additive) subgroup of $\mathbb{C}$ generated by all $\omega_i(z_j)$, for $j = 1, \ldots, r$. Since ω_i is an algebra map,

$$\omega_i(z_j)\omega_i(z_\ell) = \sum_\nu a_{j\ell\nu}\omega_i(z_\nu),$$

so that M is a ring that is finitely generated as an abelian group (because $a_{ij\nu} \in \mathbb{Z}$). Hence, for each j, M is a $\mathbb{Z}[\omega_i(z_j)]$-module that is a finitely generated abelian group. If M is faithful, then Corollary 8.10(ii) will give $\omega_i(z_j)$ an algebraic integer. But $M \subseteq \mathbb{C}$, so that the product of nonzero elements is nonzero, and this implies that M is a faithful $\mathbb{Z}[\omega_i(z_j)]$-module, as desired. •

We are almost ready to complete the proof of Burnside's theorem.

Proposition 8.152. *If $(n_i, h_j) = 1$ for some i, j, then either $|\chi_i(g_j)| = n_i$ or $\chi_i(g_j) = 0$.*

Proof. By hypothesis, there are integers s and t in $\mathbb{Z}$ with $sn_i + th_j = 1$, so that, for $g_j \in C_j$, we have

$$\frac{\chi_i(g_j)}{n_i} = s\chi_i(g_j) + \frac{th_j\chi_i(g_j)}{n_i}.$$

Hence, Proposition 8.151(iii) gives $\chi_i(g_j)/n_i$ an algebraic integer, and so $|\chi_i(g_j)| \leq n_i$, by Proposition 8.138(i); thus, it suffices to show that if $|\chi_i(g_j)/n_i| < 1$, then $\chi_i(g_j) = 0$.

Let $m(x) \in \mathbb{Z}[x]$ be the minimum polynomial of $\alpha = \chi_i(g_j)/n_i$; that is, $m(x)$ is the monic polynomial in $\mathbb{Z}[x]$ of least degree having α as a root. We proved, in Corollary 6.29, that $m(x)$ is irreducible in $\mathbb{Q}[x]$. If α' is a root of $m(x)$, then Proposition 4.13 shows that $\alpha' = \sigma(\alpha)$ for some $\sigma \in \mathrm{Gal}(E/\mathbb{Q})$, where $E/\mathbb{Q}$ is the splitting field of $m(x)(x^{|G|} - 1)$. But

$$\alpha = \frac{1}{n_i}\left(\varepsilon_1 + \cdots + \varepsilon_{n_i}\right),$$

where the ε's are $|G|$th roots of unity, and so $\alpha' = \sigma(\alpha)$ is also such a sum. It follows that $|\alpha'| \leq 1$ [as in the proof of Proposition 8.138(i)]. Therefore, if $N(\alpha)$ is the norm of α (which is, by definition, the absolute value of the product of all the roots of $m(x)$), then $N(\alpha) < 1$ (for we are assuming that $|\alpha| < 1$). But $N(\alpha)$ is the absolute value of the constant term of $m(x)$, which is an integer. Therefore, $N(\alpha) = 0$, hence $\alpha = 0$, and so $\chi_i(g_j) = 0$, as claimed. •

At last, we can prove the hypothesis of Proposition 8.116, stated at the beginning of the previous section.

Theorem 8.153. *If G is a nonabelian finite simple group, then* $\{1\}$ *is the only conjugacy class whose size is a prime power. Therefore, Burnside's theorem is true*: *every group of order $p^m q^n$, where p and q are primes, is solvable.*

Proof. Assume, on the contrary, that $h_j = p^e > 1$ for some j. By Exercise 8.62(ii) on page 633, for all i, we have

$$Z(G/\ker \chi_i) = \{g \in G : |\chi_i(g)| = n_i\}.$$

Since G is simple, $\ker \chi_i = \{1\}$ for all i, and so $Z(G/\ker \chi_i) = Z(G) = \{1\}$. By Proposition 8.152, if $(n_i, h_j) = 1$, then either $|\chi_i(g_j)| = n_i$ or $\chi_i(g_j) = 0$. Of course, $\chi_1(g_j) = 1$ for all j, where χ_1 is the trivial character. If χ_i is not the trivial character, then we have just seen that the first possibility cannot occur, and so $\chi_i(g_j) = 0$. On the other hand, if $(n_i, h_j) \neq 1$, then $p \mid n_i$ (for $h_j = p^e$). Thus, for every $i \neq 1$, either $\chi_i(g_j) = 0$ or $p \mid n_i$.

Consider the orthogonality relation, Corollary 8.136(ii):

$$\sum_{i=1}^{r} n_i \chi_i(g_j) = 0.$$

Now $n_1 = 1 = \chi_1(g_j)$, while each of the other terms is either 0 or of the form $p\alpha_i$, where α_i is an algebraic integer. It follows that

$$0 = 1 + p\beta,$$

where β is an algebraic integer. This implies that the rational number $-1/p$ is an algebraic integer, hence lies in $\mathbb{Z}$, and we have the contradiction that $-1/p$ is an integer. •

Another early application of characters is a theorem of F. G. Frobenius. We begin with a discussion of doubly transitive permutation groups. Let G be a finite group and X a finite G-set. Recall that if $x \in X$, then its *orbit* is $\mathcal{O}(x) = \{gx : g \in G\}$ and its *stabilizer* is $G_x = \{g \in G : gx = x\}$. Theorem 2.98 shows that $|\mathcal{O}(x)||G_x| = |G|$. A G-set X is *transitive* if it has only one orbit: If $x, y \in X$, then there exists $g \in G$ with $y = gx$; in this case, $\mathcal{O}(x) = X$.

If X is a G-set, then there is a homomorphism $\alpha\colon G \to S_X$, namely, $g \mapsto \alpha_g$, where $\alpha_g(x) = gx$. We say that X is a ***faithful*** G-set if α is an injection; that is, if $gx = x$ for all $x \in X$, then $g = 1$. In this case, we may regard G as a subgroup of S_X acting as permutations of X.

Cayley's theorem (Theorem 2.87) shows that every group G can be regarded as a faithful transitive G-set.

Definition. A G-set X is ***doubly transitive*** if, for every pair of 2-tuples (x_1, x_2) and (y_1, y_2) in $X \times X$ with $x_1 \neq x_2$ and $y_1 \neq y_2$, there exists $g \in G$ with $y_1 = gx_1$ and $y_2 = gx_2$.[11]

We often abuse language and call a *group* G doubly transitive if there exists a doubly transitive G-set.

Note that every doubly transitive G-set X is transitive: If $x \neq y$, then (x, y) and (y, x) are 2-tuples as in the definition, and so there is $g \in G$ with $y = gx$ (and $x = gy$).

Example 8.154.

(i) If $n \geq 2$, the symmetric group S_n is doubly transitive; that is, $X = \{1, \ldots, n\}$ is a doubly transitive S_X-set.

(ii) The alternating group A_n is doubly transitive if $n \geq 4$.

(iii) Let V be a finite-dimensional vector space over $\mathbb{F}_2$, and let $X = V - \{0\}$. Then X is a doubly transitive $\mathrm{GL}(V)$-set, for every pair of distinct nonzero vectors x_1, x_2 in V must be linearly independent (see Exercise 3.69 on page 170). Since every linearly independent list can be extended to a basis, there is a basis $x_1, x_2, \ldots, x_n$ of V. Similarly, if y_1, y_2 is another pair of distinct nonzero vectors, there is a basis $y_1, y_2, \ldots, y_n$. But $\mathrm{GL}(V)$ acts transitively on the set of all bases of V, by Exercise 3.78 on page 181. Therefore, there is $g \in \mathrm{GL}(V)$ with $y_i = gx_i$ for all i, and so X is a doubly transitive $\mathrm{GL}(V)$-set. ◂

Proposition 8.155. *A G-set X is doubly transitive if and only if, for each $x \in X$, the G_x-set $X - \{x\}$ is transitive.*

Proof. Let X be a doubly transitive G-set. If $y, z \in X - \{x\}$, then (y, x) and (z, x) are 2-tuples of distinct elements of X, and so there is $g \in G$ with $z = gy$ and $x = gx$. The latter equation shows that $g \in G_x$, and so $X - \{x\}$ is a transitive G_x-set.

To prove the converse, let (x_1, x_2) and (y_1, y_2) be 2-tuples of distinct elements of X. We must find $g \in G$ with $y_1 = gx_1$ and $y_2 = gy_2$. Let us denote (gx_1, gx_2) by $g(x_1, x_2)$. There is $h \in G_{x_2}$ with $h(x_1, x_2) = (y_1, x_2)$: if $x_1 = y_1$, we may take $h = 1_X$; if $x_1 \neq y_1$, we use the hypothesis that $X - \{x_2\}$ is a transitive G_{x_2}-set. Similarly, there is $h' \in G_{y_1}$ with $h'(y_1, x_2) = (y_1, y_2)$. Therefore, $h'h(x_1, x_2) = (y_1, y_2)$, and X is a doubly transitive G-set. •

Example 8.156.

Let k be a field, let $f(x) \in k[x]$ have no repeated roots, let E/k be a splitting field, and let $G = \mathrm{Gal}(E/k)$ be the Galois group of $f(x)$. If $X = \{\alpha_1, \ldots, \alpha_n\}$ is the set of all the roots

[11] More generally, we call a G-set X k-transitive, where $1 \leq k \leq |X|$, if, for every pair of k-tuples $(x_1, \ldots, x_k)$ and $(y_1, \ldots, y_k)$ in $X \times \cdots \times X$ having distinct coordinates, there exists $g \in G$ with $y_i = gx_i$ for all $i \leq k$. It can be proved that if $k > 5$, then the only faithful k-transitive groups are the symmetric groups and the alternating groups. The five ***Mathieu groups*** are interesting sporadic simple groups that are also highly transitive: M_{22} is 3-transitive, M_{11} and M_{23} are 4-transitive, and M_{12} and M_{24} are 5-transitive.

of $f(x)$, then X is a G-set (Theorem 4.3) that is transitive if and only if $f(x)$ is irreducible (Proposition 4.13). Now $f(x)$ factors in $k(\alpha_1)[x]$:

$$f(x) = (x - \alpha_1) f_1(x).$$

The reader may show that $G_1 = \mathrm{Gal}(E/k(\alpha_1)) \le G$ is the stabilizer G_{α_1} and that $X - \{\alpha_1\}$ is a G_1-set. Thus, Proposition 8.155 shows that X is a doubly transitive G-set if and only if both $f(x)$ and $f_1(x)$ are irreducible (over $k[x]$ and $k(\alpha_1)[x]$, respectively). ◂

Recall Example 2.92(ii): If H is a subgroup of a group G and $X = G/H$ is the family of all left cosets of H in G, then G acts on G/H by $g\colon aH \mapsto gaH$. The G-set X is transitive, and the stabilizer of $aH \in G/H$ is aHa^{-1}; that is, $gaH = aH$ if and only if $a^{-1}ga \in H$ if and only if $g \in aHa^{-1}$.

Proposition 8.157. *If X is a doubly transitive G-set, then*

$$|G| = n(n-1)|G_{x,y}|,$$

where $n = |X|$ and $G_{x,y} = \{g \in G : gx = x \text{ and } gy = y\}$. Moreover, if X is a faithful G-set, then $|G_{x,y}|$ is a divisor of $(n-2)!$.

Proof. First, Theorem 2.98 gives $|G| = n|G_x|$, because X is a transitive G-set. Now $X - \{x\}$ is a transitive G_x-set, by Proposition 8.155, and so

$$|G_x| = |X - \{x\}||(G_x)_y| = (n-1)|G_{x,y}|,$$

because $(G_x)_y = G_{x,y}$. The last remark follows, in this case, from $G_{x,y}$ being a subgroup of $S_{X-\{x,y\}} \cong S_{n-2}$. •

It is now easy to give examples of groups that are not doubly transitive, for the orders of doubly transitive groups are constrained.

Definition. A transitive G-set X is called ***regular*** if only the identity element of G fixes any element of X; that is, $G_x = \{1\}$ for all $x \in X$.

For example, Cayley's theorem shows that every group G is isomorphic to a regular subgroup of S_G. The notion of regularity extends to doubly transitive groups.

Definition. A doubly transitive G-set X is ***sharply doubly transitive*** if only the identity of G fixes two elements of X; that is, $G_{x,y} = \{1\}$ for all distinct pairs $x, y \in X$.

Proposition 8.158. *The following conditions are equivalent for a faithful doubly transitive G-set X with $|X| = n$.*

(i) *X is sharply doubly transitive.*

(ii) *If (x_1, x_2) and (y_1, y_2) are 2-tuples in $X \times X$ with $x_1 \ne x_2$ and $y_1 \ne y_2$, then there is a unique $g \in G$ with $y_1 = gx_1$ and $y_2 = gy_2$.*

(iii) $|G| = n(n-1)$.

(iv) $G_{x,y} = \{1\}$ *for all distinct* $x, y \in X$.

(v) *For every* $x \in X$, *the* G_x*-set* $X - \{x\}$ *is regular.*

Proof. All the implications are routine. •

Example 8.159.
(i) S_3 and A_4 are sharply doubly transitive groups.

(ii) The affine group $\mathrm{Aff}(1, \mathbb{R})$ was defined in Exercise 2.46 on page 80; it consists of all the functions $f : \mathbb{R} \to \mathbb{R}$ of the form $f(x) = ax + b$ with $a \neq 0$ under composition, and it is isomorphic to the subgroup of $\mathrm{GL}(2, \mathbb{R})$ consisting of all matrices of the form $\left[\begin{smallmatrix} a & b \\ 0 & 1 \end{smallmatrix}\right]$. It is plain that we can define $\mathrm{Aff}(1, k)$ for any field k in a similar way. In particular, if k is the finite field $\mathbb{F}_q$, then the affine group $\mathrm{Aff}(1, \mathbb{F}_q)$ is finite, and of order $q(q-1)$. The reader may check that $\mathbb{F}_q$ is a sharply doubly transitive $\mathrm{Aff}(1, \mathbb{F}_q)$-set. ◄

Notation. If G is a group, then $G^\# = \{g \in G : g \neq 1\}$.

By Cayley's theorem, every group is regular. We now consider transitive groups G such that each $g \in G^\#$ has at most one fixed point. In case every $g \in G^\#$ has no fixed points, then we say that the action of G is ***fixed point free***. J. G. Thompson proved that if a finite group H has a fixed point free automorphism α of prime order (that is, the action of the group $G = \langle \alpha \rangle$ on $H^\#$ is fixed point free), then H is nilpotent (see Robinson, *A Course in the Theory of Groups*, pages 306–307). Thus, let us consider such actions in which there is some $g \in G^\#$ that has a fixed point; that is, the action of G is not regular.

Definition. A finite group G is a ***Frobenius group*** if there exists a transitive G-set X such that

(i) every $g \in G^\#$ has at most one fixed point;

(ii) there is some $g \in G^\#$ that does have a fixed point.

If $x \in X$, we call G_x a ***Frobenius complement*** of G.

Note that condition (i) implies that the G-set X in the definition is necessarily faithful. Let us rephrase the two conditions: (i) that every $g \in G^\#$ has at most one fixed point says that $G_{x,y} = \{1\}$; (ii) that there is some $g \in G^\#$ that does have a fixed point says that $G_x \neq \{1\}$.

Example 8.160.
(i) The symmetric group S_3 is a Frobenius group: $X = \{1, 2, 3\}$ is a faithful transitive S_3-set; no $\alpha \in (S_3)^\#$ fixes two elements; each transposition $(i\ \ j)$ fixes one element. The cyclic subgroups $\langle (i\ \ j) \rangle$ are Frobenius complements (so Frobenius complements need not be unique). A permutation $\beta \in S_3$ has no fixed points if and only if β is a

3-cycle. We are going to prove that, in every Frobenius group, 1 together with all those elements having no fixed points comprise a normal subgroup.

(ii) The example of S_3 in part (i) can be generalized. Let X be a G-set, with at least three elements, which is a sharply doubly transitive G-set. Then X is transitive, $G_{x,y} = \{1\}$, and $G_x \neq \{1\}$ (for if $x, y, z \in X$ are distinct, there exists $g \in G$ with $x = gx$ and $z = gy$). Therefore, every sharply doubly transitive group G is a Frobenius group. ◂

Proposition 8.161. *A finite group G is a Frobenius group if and only if it contains a proper nontrivial subgroup H such that $H \cap gHg^{-1} = \{1\}$ for all $g \notin H$.*

Proof. Let X be a G-set as in the definition of Frobenius group. Choose $x \in X$, and define $H = G_x$. Now H is a proper subgroup of G, for transitivity does not permit $gx = x$ for all $g \in G$. To see that H is nontrivial, choose $g \in G^\#$ having a fixed point; say, $gy = y$. If $y = x$, then $g \in G_x = H$. If $y \neq x$, then transitivity provides $h \in G$ with $hy = x$, and Exercise 2.99 on page 114 gives $H = G_x = hG_yh^{-1} \neq \{1\}$. If $g \notin H$, then $gx \neq x$. Now $g(G_x)g^{-1} = G_{gx}$. Hence, if $h \in H \cap gHg^{-1} = G_x \cap G_{gx}$, then h fixes x and gx; that is, $h \in G_{x,y} = \{1\}$.

For the converse, we take X to be the G-set G/H of all left cosets of H in G, where $g \colon aH \mapsto gaH$ for all $g \in G$. We remarked earlier that X is a transitive G-set and that the stabilizer of $aH \in G/H$ is the subgroup aHa^{-1} of G. Since $H \neq \{1\}$, we see that $G_{aH} \neq \{1\}$. Finally, if $aH \neq bH$, then

$$G_{aH,bH} = G_{aH} \cap G_{bH} = aHa^{-1} \cap bHb^{-1} = a\big(H \cap a^{-1}bHb^{-1}a\big)a^{-1} = \{1\},$$

because $a^{-1}b \notin H$. Therefore, G is a Frobenius group. •

The significance of this last proposition is that it translates the definition of Frobenius group from the language of G-sets into the language of abstract groups.

Definition. If X is a G-set, define its ***Frobenius kernel*** to be the subset

$$N = \{1\} \cup \{g \in G : g \text{ has no fixed points}\}.$$

When X is transitive, we can describe N in terms of a stabilizer G_x. If $a \notin N^\#$, then there is some $y \in X$ with $ay = y$. Since G acts transitively, there is $g \in G$ with $gx = y$, and $a \in G_y = gG_xg^{-1}$. Hence, $a \in \bigcup_{g \in G} gG_xg^{-1}$. For the reverse inclusion, if $a \in \bigcup_{g \in G} gG_xg^{-1}$, then $a \in gG_xg^{-1} = G_{gx}$ for some $g \in G$, and so a has a fixed point; that is, $a \notin N$. We have proved that

$$N = \{1\} \cup \Big(G - \Big(\bigcup_{g \in G} gG_xg^{-1}\Big)\Big).$$

Exercise 5.32 on page 278 shows that if G_x is a proper subgroup of G, then $G \neq \bigcup_{g \in G} gG_xg^{-1}$, and so $N \neq \{1\}$ in this case.

Proposition 8.162. *If G is a Frobenius group with Frobenius complement H and Frobenius kernel N, then $|N| = [G : H]$.*

Proof. By Proposition 8.161, there is a disjoint union

$$G = \{1\} \cup \Big(\bigcup_{g \in G} gH^{\#}g^{-1} \Big) \cup N^{\#}.$$

Note that $N_G(H) = H$: If $g \notin H$, then $H \cap gHg^{-1} = \{1\}$, and so $gHg^{-1} \neq H$. Hence, the number of conjugates of H is $[G : N_G(H)] = [G : H]$ (Proposition 2.101). Therefore, $|\bigcup_{g \in G} gH^{\#}g^{-1}| = [G : H](|H| - 1)$, and so

$$|N| = |N^{\#}| + 1 = |G| - ([G : H](|H| - 1)) = [G : H]. \quad \bullet$$

The Frobenius kernel may not be a subgroup of G. It is very easy to check that if $g \in N$, then $g^{-1} \in N$ and $aga^{-1} \in N$ for every $a \in G$; the difficulty is in proving that N is closed under multiplication. For example, if $V = k^n$ is the vector space of all $n \times 1$ column vectors over a field k, then $V^{\#}$, the set of nonzero vectors in V, is a faithful transitive $\mathrm{GL}(V)$-set. Now $A \in \mathrm{GL}(V)$ has a fixed point if and only if there is some $v \in V^{\#}$ with $Av = v$; that is, A has a fixed point if and only if 1 is an eigenvalue of A. Thus, the Frobenius kernel now consists of the identity matrix together with all linear transformations which do not have 1 as an eigenvalue. Let $|k| \geq 4$, and let α be a nonzero element of k with $\alpha^2 \neq 1$. Then $A = \begin{bmatrix} \alpha & 0 \\ 0 & \alpha \end{bmatrix}$ and $B = \begin{bmatrix} \alpha^{-1} & 0 \\ 0 & \alpha \end{bmatrix}$ lie in N, but their product $AB = \begin{bmatrix} 1 & 0 \\ 0 & \alpha^2 \end{bmatrix}$ does not lie in N. However, if G is a Frobenius group, then N is a subgroup; the only known proof of this fact uses characters.

We have already remarked that if ψ is a character on a subgroup H of a group G, then the restriction $(\psi\!\uparrow^G)_H$ need not equal ψ. The next proof shows that irreducible characters of a Frobenius complement do extend to irreducible characters of G.

Lemma 8.163. *Let G be a Frobenius group with Frobenius complement H and Frobenius kernel N. For every irreducible character ψ on H other than the trivial character ψ_1, define the generalized character*

$$\varphi = \psi - d\psi_1,$$

where $d = \psi(1)$. Then $\psi^* = \varphi\!\uparrow^G + d\chi_1$ is an irreducible character on G, and $\psi^*_H = \psi$; that is, $\psi^*(h) = \psi(h)$ for all $h \in H$.

Proof. Note first that $\varphi(1) = 0$. We claim that the induced generalized character $\varphi\!\uparrow^G$ satisfies the equation

$$(\varphi\!\uparrow^G)_H = \varphi.$$

If $t_1 = 1, \ldots, t_n$ is a transversal of H in G, then for $g \in G$, the matrix of $\varphi\!\uparrow^G(g)$ on page 627 has the blocks $\dot{B}(t_i^{-1}gt_i)$ on its diagonal, where $\dot{B}(t_i^{-1}gt_i) = 0$ if $t_i^{-1}gt_i \notin H$ (this is just the matrix version of Theorem 8.142). If $h \in H$, then $t_i^{-1}ht_i \notin H$ for all $i \neq 1$, and so $\dot{B}(t_i^{-1}ht_i) = 0$. Therefore, there is only one nonzero diagonal block, and

$$\mathrm{tr}(\varphi\!\uparrow^G(h)) = \mathrm{tr}(B(h));$$

that is,

$$\varphi{\uparrow}^G(h) = \varphi(h).$$

We have just seen that $\varphi{\uparrow}^G$ is a generalized character on G such that $(\varphi{\uparrow}^G)_H = \varphi$. By Frobenius reciprocity (Theorem 8.145),

$$(\varphi{\uparrow}^G, \varphi{\uparrow}^G)_G = (\varphi, (\varphi{\uparrow}^G)_H)_H = (\varphi, \varphi)_H.$$

But $\varphi = \psi - d\psi_1$, so that orthogonality of ψ and ψ_1 gives

$$(\varphi, \varphi)_H = 1 + d^2.$$

Similarly,

$$(\varphi{\uparrow}^G, \chi_1)_G = (\varphi, \psi_1)_H = -d,$$

where χ_1 is the trivial character on G. Define

$$\psi^* = \varphi{\uparrow}^G + d\chi_1.$$

Now ψ^* is a generalized character on G, and

$$\begin{aligned}(\psi^*, \psi^*)_G &= (\varphi{\uparrow}^G, \varphi{\uparrow}^G)_G + 2d(\varphi{\uparrow}^G, \chi_1)_G + d^2\\ &= 1 + d^2 - 2d^2 + d^2 = 1.\end{aligned}$$

We have

$$(\psi^*)_H = (\varphi{\uparrow}^G)_H + d(\chi_1)_H = \varphi + d\psi_1 = (\psi - d\psi_1) + d\psi_1 = \psi.$$

Since $\psi^*(1) = \psi(1) > 0$, Corollary 8.130 says that ψ^* is an irreducible character on G. •

Theorem 8.164 (Frobenius). *Let G be a Frobenius group with Frobenius complement H and Frobenius kernel N. Then N is a normal subgroup of G, $N \cap H = \{1\}$, and $NH = G$.*

Remark. A group G having a subgroup Q and a normal subgroup K such that $K \cap Q = \{1\}$ and $KQ = G$ is called a *semidirect product*. We will discuss such groups in Chapter 10. ◂

Proof. For every irreducible character ψ on H other than the trivial character ψ_1, define the generalized character $\varphi = \psi - d\psi_1$, where $d = \psi(1)$. By the lemma, $\psi^* = \varphi{\uparrow}^G + d\chi_1$ is an irreducible character on G. Define

$$N^* = \bigcap_{\psi \neq \psi_1} \ker \psi^*.$$

Of course, N^* is a normal subgroup of G.

By Lemma 8.163, $\psi^*(h) = \psi(h)$ for all $h \in H$; in particular, if $h = 1$, we have

$$\psi^*(1) = \psi(1) = d. \tag{5}$$

If $g \in N^\#$, then for all $a \in G$, we have $g \notin aHa^{-1}$ (for g has no fixed points), and so $\dot{\varphi}(aga^{-1}) = 0$. The induced character formula, Theorem 8.142, now gives $\varphi\!\upharpoonright^G(g) = 0$. Hence, if $g \in N^\#$, then Eq. (5) gives

$$\psi^*(g) = \varphi\!\upharpoonright^G(g) + d\chi_1(g) = d.$$

We conclude that if $g \in N$, then

$$\psi^*(g) = d = \psi^*(1);$$

that is, $g \in \ker \psi^*$. Therefore,

$$N \subseteq N^*.$$

The reverse inclusion will arise from a counting argument.

Let $h \in H \cap N^*$. Since $h \in H$, Lemma 8.163 gives $\psi^*(h) = \psi(h)$. On the other hand, since $h \in N^*$, we have $\psi^*(h) = \psi^*(1) = d$. Therefore, $\psi(h) = \psi^*(h) = d = \psi(1)$, so that $h \in \ker \psi$ for every irreducible character ψ on H. Consider the regular character, afforded by the regular representation ρ on H: $\chi_\rho = \sum_i n_i \psi_i$. Now $\chi_\rho(h) = \sum_i n_i \psi_i(h) \neq 0$, so that Example 8.125(ii) gives $h = 1$. Thus,

$$H \cap N^* = \{1\}.$$

Next, $|G| = |H|[G : H] = |H||N|$, by Proposition 8.162. Note that HN^* is a subgroup of G, because $N^* \lhd G$. Now $|HN^*||H \cap N^*| = |H||N^*|$, by the second isomorphism theorem; since $H \cap N^* = \{1\}$, we have $|H||N| = |G| \geq |HN^*| = |H||N^*|$. Hence, $|N| \geq |N^*|$. But $|N| \leq |N^*|$, because $N \subseteq N^*$, and so $N = N^*$. Therefore, $N \lhd G$, $H \cap N = \{1\}$, and $HN = G$. •

Much more can be said about the structure of Frobenius groups. Every Sylow subgroup of a Frobenius complement is either cyclic or generalized quaternion (see Huppert, *Endliche Gruppen* I, page 502), and it is a consequence of J. G. Thompson's theorem on fixed-point-free automorphisms that every Frobenius kernel is nilpotent; that is, N is the direct product of its Sylow subgroups. The reader is referred to Curtis–Reiner, *Representation Theory of Finite Groups and Associative Algebras*, pages 242–246, or Feit, *Characters of Finite Groups*, pages 133–139.

EXERCISES

8.68 Prove that the affine group $\text{Aff}(1, \mathbb{F}_q)$ in Example 8.159(ii) is sharply doubly transitive.

8.69 If $H \leq G$ and the family of left cosets G/H is a G-set via the representation on cosets, prove that G/H is a faithful G-set if and only if $\bigcap_{a \in G} aHa^{-1} = \{1\}$. Give an example in which G/H is not a faithful G-set.

8.70 Prove that every Sylow subgroup of $\text{SL}(2, \mathbb{F}_5)$ is either cyclic or quaternion.

8.71 A subset A of a group G is a **T.I. set** (or a ***trivial intersection set***) if $A \subseteq N_G(A)$ and $A \cap gAg^{-1} \subseteq \{1\}$ for all $g \notin N_G(A)$.

(i) Prove that a Frobenius complement H in a Frobenius group G is a T. I. set.

(ii) Let A be a T. I. set in a finite group G, and let $N = N_G(A)$. If α be a class function vanishing on $N - A$ and β is a class function on N vanishing on $\big(\bigcup_{g \in G}(A^g \cap N)\big) - A$, prove, for all $g \in N^\#$, that $\alpha{\uparrow}^G(g) = \alpha(g)$ and $\beta{\uparrow}^G(g) = \beta(g)$.

Hint. See the proofs of Lemma 8.164 and Theorem 8.163.

(iii) If $\alpha(1) = 0$, prove that $(\alpha, \beta)_N = (\alpha{\uparrow}^G, \beta{\uparrow}^G)_G$.

(iv) Let H be a *self-normalizing* subgroup of a finite group G; that is, $H = N_G(H)$. If H is a T. I. set, prove that there is a normal subgroup K of G with $K \cap H = \{1\}$ and $KH = G$.

Hint. See Feit, *Characters of Finite Groups*, page 124.

8.72 Prove that there are no nonabelian simple groups of order n, where $60 < n \leq 100$.
Hint. By Burnside's theorem, the only candidates for n in the given range are 66, 70, 78, 84, and 90, and 90 was eliminated in Exercise 5.29(ii) on page 278.

8.73 Prove that there are no nonabelian simple groups of order n, where $101 \leq n < 168$. We remark that $\text{PSL}(2, \mathbb{F}_7)$ is a simple group of order 168, and it is the unique such group, to isomorphism. With Proposition 5.41, Corollary 5.68, and Exercise 8.72, we see that A_5 is the only nonabelian simple group of order strictly less than 168.
Hint. By Burnside's theorem, the only candidates for n in the given range are 102, 105, 110, 120, 126, 130, 132, 138, 140, 150, 154, 154, 156, and 165. Use Exercise 2.98 on page 114 and Exercises 5.30 and 5.31 on page 278.

9

Advanced Linear Algebra

This chapter begins with the study of modules over PIDs, including characterizations of their projective, injective, and flat modules. Our emphasis, however, is on finitely generated modules, because the generalization of the Fundamental Theorem of Finite Abelian Groups, when applied to $k[x]$-modules, yields the rational and Jordan canonical forms for matrices. The Smith normal form is also discussed, for it can be used to compute the invariants of a matrix. We then consider bilinear and quadratic forms on a vector space, which lead to symplectic and orthogonal groups. Multilinear algebra is the next step, leading to tensor algebras, exterior algebras, and determinants. We end with an introduction to Lie algebras, which can be viewed as a way of dealing with a family of linear transformations instead of with individual ones.

9.1 Modules over PIDs

The structure theorems for finite abelian groups will now be generalized to modules over PIDs. As we have just said, this is not mere generalization for its own sake, for the module version will yield canonical forms for matrices. Not only do the theorems generalize, but the proofs of the theorems generalize as well, as we shall see.

Definition. Let M be an R-module. If $m \in M$, then its ***order ideal*** (or ***annihilator***) is

$$\operatorname{ann}(m) = \{r \in R : rm = 0\}.$$

We say that m has ***finite order*** (or is a *torsion*[1] *element*) if $\operatorname{ann}(m) \neq \{0\}$; otherwise, m has ***infinite order***.

When a commutative ring R is regarded as a module over itself, its identity 1 has infinite order, for $\operatorname{ann}(1) = \{0\}$.

[1] The etymology of the word *torsion* is given on page 267.

Order ideals generalize the group-theoretic notion of the order of an element. Recall that if G is an additive abelian group, then an element $g \in G$ has finite order if $ng = 0$ for some positive integer n, while g has order d if d is the smallest positive integer with $dg = 0$. On the other hand, $\text{ann}(g)$ is an ideal in $\mathbb{Z}$ and, as any nonzero ideal in $\mathbb{Z}$, it is generated by the smallest positive integer in it. Thus, the order ideal $\text{ann}(g) = (d)$, the principal ideal generated by the order d of g. In Proposition 7.12, we proved that if $M = \langle m \rangle$ is a cyclic R-module, where R is any commutative ring, then $M \cong R/I$. The ideal I in this corollary is $\ker \varphi$, where $\varphi\colon R \to M$ is the map $r \mapsto rm$, so that $I = \text{ann}(m)$, and

$$\langle m \rangle \cong R/\,\text{ann}(m).$$

Definition. If M is an R-module, where R is a domain, then its ***torsion submodule***[2] tM is defined by

$$tM = \{m \in M : m \text{ has finite order}\}.$$

Proposition 9.1. *If R is a domain and M is an R-module, then tM is a submodule of M.*

Proof. If $m, m' \in tM$, then there are nonzero elements $r, r' \in R$ with $rm = 0$ and $r'm' = 0$. Clearly, $rr'(m + m') = 0$. Since R is a domain, $rr' \neq 0$, and so $\text{ann}(m + m') \neq \{0\}$; therefore, $m + m' \in tM$.

If $s \in R$, then $sm \in tM$, for $r \in \text{ann}(sm)$ because $rsm = 0$. •

This proposition can be false if R is not a domain. For example, let $R = \mathbb{I}_6$. In $M = \mathbb{I}_6$, both [3] and [4] have finite order, for $[2] \in \text{ann}([3])$ and $[3] \in \text{ann}([4])$. On the other hand, $[3] + [4] = [1]$, and [1] has infinite order in M, for $\text{ann}([1]) = \{0\}$.

For the remainder of this section, R will be a domain (indeed, it will soon be restricted even further).

Definition. If R is a domain and M is an R-module, then M is ***torsion*** if $tM = M$, while M is ***torsion-free*** if $tM = \{0\}$.

Proposition 9.2. *Let M and M' be R-modules, where R is a domain.*

(i) *M/tM is torsion-free.*

(ii) *If $M \cong M'$, then $tM \cong tM'$ and $M/tM \cong M'/tM'$.*

Proof. (i) Assume that $m + tM \neq 0$ in M/tM; that is, m has infinite order. If $m + tM$ has finite order, then there is some $r \in R$ with $r \neq 0$ such that $0 = r(m + tM) = rm + tM$; that is, $rm \in tM$. Thus, there is $s \in R$ with $s \neq 0$ and with $0 = s(rm) = (sr)m$. But $sr \neq 0$, since R is a domain, and so $\text{ann}(m) \neq \{0\}$; this contradicts m having infinite order.

[2]There is a generalization of the torsion submodule, called the ***singular submodule***, which is defined for left R-modules over any not necessarily commutative ring. See Dauns, *Modules and Rings*, pages 231–238.

(ii) If $\varphi\colon M \to M'$ is an isomorphism, then $\varphi(tM) \subseteq tM'$, for if $rm = 0$ with $r \neq 0$, then $r\varphi(m) = \varphi(rm) = 0$ (this is true for any R-homomorphism); hence, $\varphi|tM\colon tM \to tM'$ is an isomorphism (with inverse $\varphi^{-1}|tM'$). For the second statement, the map $\overline{\varphi}\colon M/tM \to M'/tM'$, defined by $\overline{\varphi}\colon m + tM \mapsto \varphi(m) + tM'$, is easily seen to be an isomorphism. •

Here is a fancy proof of Proposition 9.2. There is a functor $t\colon {}_R\mathbf{Mod} \to {}_R\mathbf{Mod}$ defined on modules by $M \mapsto tM$ and on morphisms by $\varphi \mapsto \varphi|tM$. That $tM \cong tM'$ follows from the fact that every functor preserves equivalences.

A non-noetherian commutative ring R, by its very definition, has an ideal that is not finitely generated. Now R, viewed as a module over itself, is finitely generated; indeed, it is cyclic (with generator 1). Thus, it is possible that a submodule of a finitely generated module need not, itself, be finitely generated. This cannot happen when R is a PID; in fact, we have proved, in Proposition 7.23(ii), that if R is a PID, then every submodule S of a finitely generated R-module is itself finitely generated; indeed, if M can be generated by n elements, then S can be generated by n or fewer elements.

Theorem 9.3. *If R is a* PID*, then every finitely generated torsion-free R-module M is free.*

Proof. We prove the theorem by induction on n, where $M = \langle v_1, \ldots, v_n\rangle$.

If $n = 1$, then M is cyclic; hence, $M = \langle v_1\rangle \cong R/\operatorname{ann}(v_1)$. Since M is torsion-free, $\operatorname{ann}(v_1) = \{0\}$, so that $M \cong R$, and hence M is free.

For the inductive step, let $M = \langle v_1, \ldots, v_{n+1}\rangle$ and define

$$S = \{m \in M : \text{there is } r \in R, r \neq 0, \text{ with } rm \in \langle v_{n+1}\rangle\};$$

it is easy to check that S is a submodule of M. Now M/S is torsion-free: If $x \in M$, $x \notin S$, and $r(x+S) = 0$, then $rx \in S$; hence, there is $r' \in R$ with $r' \neq 0$ and $rr'x \in \langle v_{n+1}\rangle$. Since $rr' \neq 0$, we have $x \in S$, a contradiction. Plainly, M/S can be generated by n elements, namely, $v_1 + S, \ldots, v_n + S$, and so M/S is free, by the inductive hypothesis. Since free modules are projective, Proposition 7.54 gives

$$M \cong S \oplus (M/S).$$

Thus, the proof will be completed once we prove that $S \cong R$.

If $x \in S$, then there is some nonzero $r \in R$ with $rx \in \langle v_{n+1}\rangle$; that is, there is $a \in R$ with $rx = av_{n+1}$. Define $\varphi\colon S \to Q = \operatorname{Frac}(R)$, the fraction field of R, by $\varphi\colon x \mapsto a/r$. It is a straightforward calculation, left to the reader, that φ is a (well-defined) injective R-map. If $D = \operatorname{im}\varphi$, then D is a finitely generated submodule of Q.

The proof will be complete if we can prove that every finitely generated submodule D of Q is cyclic. Now

$$D = \langle b_1/c_1, \ldots, b_m/c_m\rangle,$$

where $b_i, c_i \in R$. Let $c = \prod_i c_i$, and define $f\colon D \to R$ by $f\colon d \mapsto cd$ for all $d \in D$ (it is plain that f has values in R, for multiplication by c clears all denominators). Since D is torsion-free, f is an injective R-map, and so D is isomorphic to a submodule of R; that is, D is isomorphic to an ideal of R. Since R is a PID, every nonzero ideal in R is isomorphic to R; hence, $S \cong \operatorname{im}\varphi = D \cong R$. •

Corollary 9.4. *If R is a* PID, *then every submodule S of a finitely generated free R-module F is itself free, and* $\text{rank}(S) \leq \text{rank}(F)$. *In particular, every finitely generated projective R-module P is free.*

Proof. By Proposition 7.23(ii), the submodule S can be generated by n or fewer elements, where $n = \text{rank}(F)$. Now F is torsion-free, and hence S is torsion-free. Theorem 9.3 now applies to give S free.

The second statement follows from Theorem 7.56: the characterization of projective modules as direct summands of free modules. Since P is finitely generated, there is a finitely generated free module F and a surjection $q\colon F \to P$; since P is projective, there is a map $j\colon P \to F$ with $qj = 1_P$. Thus, j restricts to an isomorphism of P with a submodule of F, which is free, by the first part of the proof. •

Remark. Both statements in the corollary are true without the finiteness hypothesis, and we shall soon prove them. ◂

Corollary 9.5.

(i) *If R is a* PID, *then every finitely generated R-module M is a direct sum*

$$M = tM \oplus F,$$

where F is a finitely generated free R-module.

(ii) *If M and M' are finitely generated R-modules, where R is a PID, then $M \cong M'$ if and only if $tM \cong tM'$ and* $\text{rank}(M/tM) = \text{rank}(M'/tM')$.

Proof. (i) The quotient module M/tM is finitely generated, because M is finitely generated, and it is torsion-free, by Proposition 9.2(i). Therefore, M/tM is free, by Theorem 9.3, and hence M/tM is projective. Finally, $M \cong tM \oplus (M/tM)$, by Corollary 7.55 on page 476.

(ii) By Proposition 9.2(ii), if $M \cong M'$, then $tM \cong tM'$ and $M/tM \cong M'/tM'$. Since M/tM is finitely generated torsion-free, it is a free module, as is M'/tM', and these are isomorphic if they have the same rank.

Conversely, since $M \cong tM \oplus (M/tM)$ and $M' \cong tM' \oplus (M'/tM')$, Proposition 7.30 assembles the isomorphisms on each summand into an isomorphism $M \to M'$. •

Remark. This corollary requires the finitely generated hypothesis. There exist abelian groups G whose torsion subgroup tG is not a direct summand of G [see Exercise 9.1(iii) on page 663]. ◂

We can now characterize flat modules over a PID.

Corollary 9.6. *If R is a* PID*, then an R-module M is flat if and only if it is torsion-free.*

Proof. By Theorem 9.3, every finitely generated torsion-free R-module is free, and so it is flat, by Lemma 8.98. By Lemma 8.97, M itself is flat.

Conversely, if M is not torsion-free, then it contains a nonzero element m of finite order, say, (r). If $i\colon R \to \mathrm{Frac}(R)$ is the inclusion, then $m \otimes 1 \in \ker(1_M \otimes i)$, for in $M \otimes_R \mathrm{Frac}(R)$, we have

$$m \otimes 1 = m \otimes \frac{r}{r} = rm \otimes \frac{1}{r} = 0.$$

On the other hand, $m \otimes 1 \neq 0$ in $M \otimes_R R$, for the map $m \otimes 1 \mapsto m$ is an isomorphism $M \otimes_R R \to M$, by Proposition 8.86. Therefore, M is not flat. •

Before continuing the saga of finitely generated modules, we pause to prove an important result: the generalization of Corollary 9.4, in which we no longer assume that free modules F are finitely generated. We begin with a second proof of the finitely generated case that will then be generalized.

Proposition 9.7. *If R is a* PID*, then every submodule H of a finitely generated free R-module F is itself free, and* $\mathrm{rank}(H) \le \mathrm{rank}(F)$.

Proof. The proof is by induction on $n = \mathrm{rank}(F)$. If $n = 1$, then $F \cong R$. Thus, H is isomorphic to an ideal in R; but all ideals are principal, and hence are isomorphic to $\{0\}$ or R. Therefore, H is a free module of rank ≤ 1.

Let us now prove the inductive step. If $\{x_1, \ldots, x_{n+1}\}$ is a basis of F, define $F' = \langle x_1, \ldots, x_n\rangle$, and let $H' = H \cap F'$. By induction, H' is a free module of rank $\le n$. Now

$$H/H' = H/(H \cap F') \cong (H + F')/F' \subseteq F/F' \cong R.$$

By the base step, either $H/H' = \{0\}$ or $H/H' \cong R$. In the first case, $H = H'$, and we are done. In the second case, Corollary 7.55 gives $H = H' \oplus \langle h\rangle$ for some $h \in H$, where $\langle h\rangle \cong R$, and so H is free abelian of rank $\le n + 1$. •

We now remove the finiteness hypothesis.

Theorem 9.8. *If R is a* PID*, then every submodule H of a free R-module F is itself free, and* $\mathrm{rank}(H) \le \mathrm{rank}(F)$. *In particular, every projective R-module H is free.*

Proof. We are going to use the statement, equivalent to the axiom of choice and to Zorn's lemma (see the Appendix), that every set can be well-ordered. In particular, we may assume that $\{x_k : k \in K\}$ is a basis of F having a well-ordered index set K.

For each $k \in K$, define

$$F'_k = \langle x_j : j \prec k\rangle \quad \text{and} \quad F_k = \langle x_j : j \preceq k\rangle = F'_k \oplus \langle x_k\rangle;$$

note that $F = \bigcup_k F_k$. Define

$$H'_k = H \cap F'_k \quad \text{and} \quad H_k = H \cap F_k.$$

Now $H'_k = H \cap F'_k = H_k \cap F'_k$, so that

$$\begin{aligned} H_k/H'_k &= H_k/(H_k \cap F'_k) \\ &\cong (H_k + F'_k)/F'_k \subseteq F_k/F'_k \cong R. \end{aligned}$$

By Corollary 7.55, either $H_k = H'_k$ or $H_k = H'_k \oplus \langle h_k \rangle$, where $h_k \in H_k \subseteq H$ and $\langle h_k \rangle \cong R$. We claim that H is a free R-module with basis the set of all h_k. It will then follow that $\operatorname{rank}(H) \le \operatorname{rank}(F)$.

Since $F = \bigcup F_k$, each $f \in F$ lies in some F_k; since K is well-ordered, there is a smallest index $k \in K$ with $f \in F_k$, and we denote this smallest index by $\mu(f)$. In particular, if $h \in H$, then

$$\mu(h) = \text{ smallest index } k \text{ with } h \in F_k.$$

Note that if $h \in H'_k \subseteq F'_k$, then $\mu(h) \prec k$. Let H^* be the submodule of H generated by all the h_k.

Suppose that H^* is a proper submodule of H. Let j be the smallest index in

$$\{\mu(h) : h \in H \text{ and } h \notin H^*\},$$

and choose $h' \in H$ to be such an element having index j; that is, $h' \notin H^*$ and $\mu(h') = j$. Now $h' \in H \cap F_j$, because $\mu(h') = j$, and so

$$h' = a + rh_j, \text{ where } a \in H'_j \text{ and } r \in R.$$

Thus, $a = h' - rh_j \in H'_j$ and $a \notin H^*$; otherwise $h' \in H^*$ (because $h_j \in H^*$). Since $\mu(a) \prec j$, we have contradicted j being the smallest index of an element of H not in H^*. We conclude that $H^* = H$; that is, every $h \in H$ is a linear combination of h_k's.

It remains to prove that an expression of any $h \in H$ as a linear combination of h_k's is unique. By subtracting two such expressions, it suffices to prove that if

$$0 = r_1 h_{k_1} + r_2 h_{k_2} + \cdots + r_n h_{k_n},$$

then all the coefficients $r_i = 0$. Arrange the terms so that $k_1 \prec k_2 \prec \cdots \prec k_n$. If $r_n \neq 0$, then $r_n h_{k_n} \in \langle h_{k_n} \rangle \cap H'_{k_n} = \{0\}$, a contradiction. Therefore, all $r_i = 0$, and so H is a free module with basis $\{h_k : k \in K\}$. •

We return to the discussion of finitely generated modules. In light of Proposition 9.2(ii), the problem of classifying finitely generated R-modules, when R is a PID, is reduced to classifying finitely generated torsion modules. Let us say at once that these modules are precisely the generalization of finite abelian groups.

Proposition 9.9. *An abelian group G is finite if and only if it is a finitely generated torsion $\mathbb{Z}$-module.*

Proof. If G is finite, it is obviously finitely generated; moreover, Lagrange's theorem says that G is torsion.

Conversely, suppose that $G = \langle x_1, \ldots, x_n \rangle$ and there are nonzero integers d_i with $d_i x_i = 0$ for all i. It follows that each $g \in G$ can be written

$$g = m_1 x_1 + \cdots + m_n x_n,$$

where $0 \le m_i < d_i$ for all i. Therefore, $|G| \le \prod_i d_i$, and so G is finite. •

Definition. Let R be a PID and M be an R-module. If $P = (p)$ is a nonzero prime ideal in R, then M is (p)-***primary*** if, for each $m \in M$, there is $n \ge 1$ with $p^n m = 0$.

If M is any R-module, then its (p)-***primary component*** is

$$M_P = \{m \in M : p^n m = 0 \text{ for some } n \ge 1\}.$$

If we do not want to specify the prime P, we may write that a module is *primary* (instead of P-primary). It is clear that primary components are submodules.

All of the coming theorems in this section were first proved for abelian groups and, later, generalized to modules over PIDs. The translation from abelian groups to modules is straightforward, but let us see this explicitly by generalizing the primary decomposition to modules over PIDs by adapting the proof given in Chapter 5 for abelian groups. For the reader's convenience, we reproduce this proof with the finiteness hypothesis eliminated.

Theorem 9.10 (Primary Decomposition).

(i) *Every torsion abelian group G is a direct sum of its p-primary components*:

$$G = \sum_p G_p.$$

(ii) *Every torsion R-module M, where R is a* PID*, is a direct sum of its P-primary components*:

$$M = \sum_P M_P.$$

Proof. (i) Let $x \in G$ be nonzero, and let its order be d. By the fundamental theorem of arithmetic, there are distinct primes $p_1, \ldots, p_n$ and positive exponents $e_1, \ldots, e_n$ with

$$d = p_1^{e_1} \cdots p_n^{e_n}.$$

Define $r_i = d/p_i^{e_i}$, so that $p_i^{e_i} r_i = d$. It follows that $r_i x \in G_{p_i}$ for each i. But the gcd of $r_1, \ldots, r_n$ is 1, and so there are integers $s_1, \ldots, s_n$ with $1 = \sum_i s_i r_i$. Therefore,

$$x = \sum_i s_i r_i x \in \Big\langle \bigcup_p G_p \Big\rangle.$$

For each prime p, write $H_p = \left\langle \bigcup_{q \neq p} G_q \right\rangle$. By Exercise 7.79 on page 519, it suffices to prove that if

$$x \in G_p \cap H_p,$$

then $x = 0$. Since $x \in G_p$, we have $p^\ell x = 0$ for some $\ell \geq 0$; since $x \in H_p$, we have $ux = 0$, where $u = q_1^{f_1} \cdots q_n^{f_n}$, $q_i \neq p$, and $f_i \geq 1$ for all i. But p^ℓ and u are relatively prime, so there exist integers s and t with $1 = sp^\ell + tu$. Therefore,

$$x = (sp^\ell + tu)x = sp^\ell x + tux = 0.$$

(ii) We now translate the proof just given into the language of modules. If $m \in M$ is nonzero, its order ideal $\operatorname{ann}(m) = (d)$, for some $d \in R$. By unique factorization, there are irreducible elements $p_1, \ldots, p_n$, no two of which are associates, and positive exponents $e_1, \ldots, e_n$ with

$$d = p_1^{e_1} \cdots p_n^{e_n}.$$

By Proposition 6.17, $P_i = (p_i)$ is a prime ideal for each i. Define $r_i = d/p_i^{e_i}$, so that $p_i^{e_i} r_i = d$. It follows that $r_i m \in M_{P_i}$ for each i. But the gcd of the elements $r_1, \ldots, r_n$ is 1, and so there are elements $s_1, \ldots, s_n \in R$ with $1 = \sum_i s_i r_i$. Therefore,

$$m = \sum_i s_i r_i m \in \left\langle \bigcup_P M_P \right\rangle.$$

For each prime P, write $H_P = \left\langle \bigcup_{Q \neq P} G_Q \right\rangle$. By Exercise 7.79 on page 519, it suffices to prove that if

$$m \in M_P \cap H_P,$$

then $m = 0$. Since $m \in M_P$ where $P = (p)$, we have $p^\ell m = 0$ for some $\ell \geq 0$; since $m \in H_P$, we have $um = 0$, where $u = q_1^{f_1} \cdots q_n^{f_n}$, $Q_i = (q_i)$, and $f_i \geq 1$. But p^ℓ and u are relatively prime, so there exist $s, t \in R$ with $1 = sp^\ell + tu$. Therefore,

$$m = (sp^\ell + tu)m = sp^\ell m + tum = 0. \quad \bullet$$

Proposition 9.11. *Two torsion modules M and M' over a* PID *are isomorphic if and only if $M_P \cong M'_P$ for every nonzero prime ideal P.*

Proof. If $f: M \to M'$ is an R-map, then $f(M_P) \subseteq M'_P$ for every prime ideal $P = (p)$, for if $p^\ell m = 0$, then $0 = f(p^\ell m) = p^\ell f(m)$. If f is an isomorphism, then $f^{-1}: M' \to M$ is also an isomorphism. It follows that each restriction $f|M_P: M_P \to M'_P$ is an isomorphism, with inverse $f^{-1}|M'_P$. Conversely, if there are isomorphisms $f_P: M_P \to M'_P$ for all P, then there is an isomorphism $\varphi: \sum_P M_P \to \sum_P M'_P$ given by $\sum_P m_P \mapsto \sum_P f_P(m_P)$. $\bullet$

We remark that there is a fancy proof here, just as there is for Proposition 9.2. Define a "P-torsion functor" $t_P\colon {}_R\mathbf{Mod} \to {}_R\mathbf{Mod}$ on modules by $M \mapsto (tM)_P$ and on morphisms by $\varphi \mapsto \varphi|(tM)_P$. That $(tM)_P \cong (tM')_P$ follows from the fact that every functor preserves equivalences.

For the remainder of this section, we shall merely give definitions and statements of results; the reader should have no difficulty in adapting proofs of theorems about abelian groups to proofs of theorems about modules over PIDs.

Theorem 9.12 (Basis Theorem). *If R is a* PID*, then every finitely generated module M is a direct sum of cyclic modules in which each cyclic summand is either primary or is isomorphic to R.*

Proof. By Corollary 9.5, $M = tM \oplus F$, where F is finitely generated free; see Theorem 5.18 for the abelian group version of the basis theorem. •

Remark. The reader may be amused by a sophisticated proof of the basis theorem. By Corollary 9.5 and Theorem 9.10(ii), we may assume that M is P-primary for some prime ideal $P = (p)$.

There is a positive integer e with $p^e M = \{0\}$: if $M = \langle m_1, \ldots, m_n\rangle$, then $p^{e_i} m_i = 0$ for some e_i, and we choose e to be the largest of the e_i (we may assume that $e = e_n$). By Exercise 7.4 on page 440, if $J = (p^e)$, then M/JM is an R/J-module; indeed, since $JM = \{0\}$, we have M itself is an R/J-module. Now $\langle m_n\rangle \cong R/(p^e) = R/J$ is an injective R/J-module, by Proposition 7.76, and Proposition 7.64 says that the submodule $S = \langle m_n\rangle$ is a direct summand:

$$M = \langle m_n\rangle \oplus T,$$

where T is an R/J-submodule of M; a fortiori, T is an R-submodule of M [if $r \in R$ and $t \in T$, then $(r + J)t$ makes sense; define $rt = (r + J)t$]. As T can be generated by fewer than n elements, we may assume, by induction, that it is a direct sum of cyclic submodules. ◄

Corollary 9.13. *Every finitely generated abelian group is a direct sum of cyclic groups, each of prime power order or infinite.*

When are two finitely generated modules M and M' over a PID isomorphic?

Before stating the next lemma, recall that M/pM is a vector space over $R/(p)$, and we define

$$d(M) = \dim(M/pM).$$

In particular, $d(pM) = \dim(pM/p^2M)$ and, more generally,

$$d(p^n M) = \dim(p^n M/p^{n+1}M).$$

Definition. If M is a finitely generated (p)-primary R-module, where R is a PID and $P = (p)$ is a prime ideal, then

$$U_P(n, M) = d(p^n M) - d(p^{n+1}M).$$

Theorem 9.14. *If R is a* PID *and $P = (p)$ is a prime ideal in R, then any two decompositions of a finitely generated P-primary R-module M into direct sums of cyclic modules have the same number of cyclic summands of each type. More precisely, for each $n \geq 0$, the number of cyclic summands having order ideal (p^{n+1}) is $U_P(n, M)$.*

Proof. See Theorem 5.23. •

Corollary 9.15. *If M and M' are P-primary R-modules, where R is a* PID, *then $M \cong M'$ if and only if $U_P(n, M) = U_P(n, M')$ for all $n \geq 0$.*

Proof. See Corollary 5.24. •

Definition. If M is a P-primary R-module, where R is a PID, then the ***elementary divisors*** of M are the ideals (p^{n+1}), each repeated with multiplicity $U_P(n, M)$.

If M is a finitely generated torsion R-module, then its ***elementary divisors*** are the elementary divisors of all its primary components.

The next definition is motivated by Corollary 5.30: If G is a finite abelian group with elementary divisors $\{p_i^{e_{ij}}\}$, then

$$|G| = \prod_{ij} p_i^{e_{ij}}.$$

Definition. If M is a finitely generated torsion R-module, where R is a PID, then the ***order*** of M is the principal ideal generated by the product of its elementary divisors, namely, $\left(\prod_{ij} p_i^{e_{ij}}\right)$.

Example 9.16.

If k is a field, how many $k[x]$-modules are there of order $(x-1)^3(x+1)^2$? By the primary decomposition, every $k[x]$-module of order $(x-1)^3(x+1)^2$ is the direct sum of primary modules of order $(x-1)^3$ and $(x+1)^2$, respectively. There are three modules of order $(x-1)^3$, described by the elementary divisors

$$(x-1, x-1, x-1), \quad (x-1, (x-1)^2), \quad \text{and} \quad (x-1)^3;$$

there are two modules of order $(x+1)^2$, described by the elementary divisors

$$(x+1, x+1) \quad \text{and} \quad (x+1)^2.$$

Therefore, to isomorphism, there are six modules of order $(x-1)^3(x+1)^2$.

The reader has probably noticed that this argument is same as that in Example 5.26 on page 264 classifying all abelian groups of order $72 = 2^3 3^2$. ◂

Theorem 9.17 (Fundamental Theorem of Finitely Generated Modules). *If R is a* PID*, then two finitely generated R-modules are isomorphic if and only if their torsion submodules have the same elementary divisors and their free parts have the same rank.*

Proof. By Theorem 9.10(ii), $M \cong M'$ if and only if, for all primes P, the primary components M_P and M'_P are isomorphic. Corollary 9.15, Proposition 9.5(ii), and Proposition 9.11 now complete the proof. •

Here is a second type of decomposition of a finitely generated torsion module into a direct sum of cyclics that does not mention primary modules.

Proposition 9.18. *If R is a* PID*, then every finitely generated torsion R-module M is a direct sum of cyclic modules*

$$M = R/(c_1) \oplus R/(c_2) \oplus \cdots \oplus R/(c_t),$$

where $t \geq 1$ and $c_1 \mid c_2 \mid \cdots \mid c_t$.

Proof. See Proposition 5.27. •

Definition. If M is a finitely generated torsion R-module, where R is a PID, and if

$$M = R/(c_1) \oplus R/(c_2) \oplus \cdots \oplus R/(c_t),$$

where $t \geq 1$ and $c_1 \mid c_2 \mid \cdots \mid c_t$, then $(c_1), (c_2), \ldots, (c_t)$ are called the ***invariant factors*** of M.

Corollary 9.19. *If M is a finitely generated torsion module over a* PID *R, then*

$$(c_t) = \big\{r \in R : rM = \{0\}\big\},$$

where (c_t) is the last ideal occurring in the decomposition of M in Proposition 9.18.

In particular, if $R = k[x]$, where k is a field, then c_t is the polynomial of least degree for which $c_t M = \{0\}$.

Proof. For the first statement, see Corollary 5.28.

The second statement follows from the fact that every nonzero ideal in $k[x]$ is generated by the monic polynomial of least degree in it. •

Definition. If M is an R-module, then its ***exponent*** (or ***annihilator***) is the ideal

$$\operatorname{ann}(M) = \{r \in R : rM = \{0\}\}.$$

Corollary 9.19 computes the exponent of a finitely generated torsion module over a PID; it is the last invariant factor (c_t).

Corollary 9.20. *If M is a finitely generated torsion R-module, where R is a* PID*, with invariant factors $c_1, \ldots, c_t$, then the order of M is $\left(\prod_{i=1}^{t} c_i\right)$.*

Proof. See Corollary 5.30. The reader should check that the principal ideal generated by the product of the elementary divisors (which is the definition of the order of M) is equal to the principal ideal $\left(\prod_{i=1}^{t} c_i\right)$. •

Example 9.21.
We displayed the elementary divisors of $k[x]$-modules of order $(x-1)^3(x+1)^2$ in Example 9.16; here are their invariant factors.

Elementary divisors	$\leftrightarrow$	Invariant factors
$(x-1, x-1, x-1, x+1, x+1)$	$\leftrightarrow$	$x-1 \mid (x-1)(x+1) \mid (x-1)(x+1)$
$(x-1, (x-1)^2, x+1, x+1)$	$\leftrightarrow$	$(x-1)(x+1) \mid (x-1)^2(x+1)$
$((x-1)^3, x+1, x+1)$	$\leftrightarrow$	$x+1 \mid (x-1)^3(x+1)$
$(x-1, x-1, x-1, (x+1)^2)$	$\leftrightarrow$	$x-1 \mid x-1 \mid (x-1)(x+1)^2$
$(x-1, (x-1)^2, (x+1)^2)$	$\leftrightarrow$	$x-1 \mid (x-1)^2(x+1)^2$
$((x-1)^3, (x+1)^2)$	$\leftrightarrow$	$(x-1)^3(x+1)^2$ ◂

Theorem 9.22 (Invariant Factors). *If R is a* PID*, then two finitely generated R-modules are isomorphic if and only if their torsion submodules have the same invariant factors and their free parts have the same rank.*

Proof. By Corollary 9.5(i), every finitely generated R-module M is a direct sum $M = tM \oplus F$, where F is free, and $M \cong M'$ if and only if $tM \cong tM'$ and $F \cong F'$. Corollary 9.5(ii) shows that the free parts $F \cong M/tM$ and $F' \cong M'/tM'$ are isomorphic, and a straightforward generalization of Theorem 5.32 shows that the torsion submodules are isomorphic. •

The reader should now be comfortable when we say that a theorem can easily be generalized from abelian groups to modules over PID's. Consequently, we will state and prove theorems only for abelian groups, leaving the straightforward generalizations to modules to the reader.

Let us now consider modules that are not finitely generated. Recall that an abelian group D is ***divisible*** if, for each $d \in D$ and each positive integer n, there exists $d' \in D$ with $d = nd'$. Every quotient of a divisible group is divisible, as is every direct sum of divisible groups. Now Corollary 7.73 states that an abelian group D is an injective $\mathbb{Z}$-module if and only if it is divisible, so that classifying divisible abelian groups describes all injective abelian groups.

Proposition 9.23. *A torsion-free abelian group D is divisible if and only if it is a vector space over $\mathbb{Q}$.*

Proof. If D is a vector space over $\mathbb{Q}$, then it is a direct sum of copies of $\mathbb{Q}$, for every vector space has a basis. But $\mathbb{Q}$ is a divisible group, and any direct sum of divisible groups is itself divisible.

Let D be torsion-free and divisible; we must show that D admits scalar multiplication by rational numbers. Suppose that $d \in D$ and n is a positive integer. Since D is divisible, there exists $d' \in D$ with $nd' = d$ [of course, d' is a candidate for $(1/n)d$]. Note, since D is torsion-free, that d' is the unique such element: If also $nd'' = d$, then $n(d' - d'') = 0$, so that $d' - d''$ has finite order, and hence is 0. If $m/n \in \mathbb{Q}$, define $(m/n)d = md'$, where $nd' = d$. It is a routine exercise for the reader to prove that this scalar multiplication is well-defined [if $m/n = a/b$, then $(m/n)d = (a/b)d$] and that the various axioms in the definition of vector space hold. •

Definition. If G is an abelian group, then dG is the subgroup generated by all the divisible subgroups of G.

Proposition 9.24.

(i) *For any abelian group G, the subgroup dG is the unique maximal divisible subgroup of G.*

(ii) *Every abelian group G is a direct sum*

$$G = dG \oplus R,$$

where $dR = \{0\}$. Hence, $R \cong G/dG$ has no nonzero divisible subgroups.

Proof. (i) It suffices to prove that dG is divisible, for then it is obviously the largest such. If $x \in dG$, then $x = x_1 + \cdots + x_t$, where $x_i \in D_i$ and the D_i are divisible subgroups of G. If n is a positive integer, then there are $y_i \in D_i$ with $x_i = ny_i$, because D_i is divisible. Hence, $y = y_1 + \cdots + y_t \in dG$ and $x = ny$, so that dG is divisible.

(ii) Since dG is divisible, it is injective, and Proposition 7.64 gives

$$G = dG \oplus R,$$

where R is a subgroup of G. If R has a nonzero divisible subgroup D, then $R = D \oplus S$ for some subgroup S, by Proposition 7.64 on page 481. But $dG \oplus D$ is a divisible subgroup of G properly containing dG, and this contradicts part (i). •

Definition. An abelian group G is ***reduced*** if $dG = \{0\}$; that is, G has no nonzero divisible subgroups.

In Exercise 9.18 on page 665, we prove that an abelian group G is reduced if and only if $\mathrm{Hom}(\mathbb{Q}, G) = \{0\}$.

We have just shown that G/dG is always reduced. The reader should compare the roles of the maximal divisible subgroup dG of a group G with that of tG, its torsion subgroup:

G is torsion if $tG = G$ and it is torsion-free if $tG = \{0\}$; G is divisible if $dG = G$ and it is reduced if $dG = \{0\}$. There are exact sequences

$$0 \to dG \to G \to G/dG \to 0$$

and

$$0 \to tG \to G \to G/tG \to 0;$$

the first sequence always splits, but we will see, in Exercise 9.1(iii) on page 663, that the second sequence may not split.

The following group has some remarkable properties.

Definition. If p is a prime, a complex number z is a ***pth-power root of unity*** if $z^{p^n} = 1$ for some $n \geq 1$. The ***quasicyclic group*** (also called the *Prüfer group* of type p^∞) is

$$\mathbb{Z}(p^\infty) = \{\text{complex } p\text{th power roots of unity}\}.$$

Of course, if z is a pth power root of unity, say, $z^{p^n} = 1$, then z is a power of the primitive p^nth root of unity $z_n = e^{2\pi i/p^n}$. Note, for every integer $n \geq 1$, that the subgroup $\langle z_n \rangle$ is the unique subgroup of $\mathbb{Z}(p^\infty)$ of order p^n, for the polynomial $x^{p^n} - 1 \in \mathbb{C}[x]$ has at most p^n complex roots.

Proposition 9.25. *Let p be a prime.*

(i) *$\mathbb{Z}(p^\infty)$ is isomorphic to the p-primary component of $\mathbb{Q}/\mathbb{Z}$.*

(ii) *$\mathbb{Z}(p^\infty)$ is a divisible p-primary abelian group.*

(iii) *The subgroups of $\mathbb{Z}(p^\infty)$ are*

$$\{1\} \subsetneq \langle z_1 \rangle \subsetneq \langle z_2 \rangle \subsetneq \cdots \subsetneq \langle z_n \rangle \subsetneq \langle z_{n+1} \rangle \subsetneq \cdots \subsetneq \mathbb{Z}(p^\infty),$$

and so they are well-ordered by inclusion.[3]

(iv) *$\mathbb{Z}(p^\infty)$ has the* DCC *on subgroups but not the* ACC.[4]

Proof. (i) Define $\varphi\colon \sum_p \mathbb{Z}(p^\infty) \to \mathbb{Q}/\mathbb{Z}$ by $\varphi\colon (e^{2\pi i c_p/p^{n_p}}) \mapsto \sum_p c_p/p^{n_p} + \mathbb{Z}$, where $c_p \in \mathbb{Z}$. It is easy to see that φ is an injective homomorphism. The proof that φ is surjective is really contained in the proof of Theorem 5.13, but here it is again. Let $a/b \in \mathbb{Q}/\mathbb{Z}$, and write $b = \prod_p p^{n_p}$. Since the numbers b/p^{n_p} are pairwise relatively prime, there are integers m_p with $1 = \sum_p m_p(b/p^{n_p})$. Therefore, $a/b = \sum_p am_p/p^{n_p} = \varphi((am_p/p^{n_p}))$.

(ii) Since a direct summand is always a homomorphic image, $\mathbb{Z}(p^\infty)$ is a homomorphic image of the divisible group $\mathbb{Q}/\mathbb{Z}$; but every quotient of a divisible group is itself divisible.

---

[3] The group $\mathbb{Z}(p^\infty)$ is called *quasicyclic* because every proper subgroup of it is cyclic.

[4] Theorem 8.46, the Hopkins–Levitzki theorem, says that a ring with DCC must also have ACC. This result shows that the analogous result for groups is false.

(iii) Let S be a proper subgroup of $\mathbb{Z}(p^\infty)$. Since $\{z_n : n \geq 1\}$ generates $\mathbb{Z}(p^\infty)$, we may assume that $z_m \notin S$ for some m. It follows that $z_\ell \notin S$ for all $\ell > m$; otherwise $z_m = z_\ell^{p^{\ell-m}} \in S$. If $S \neq \{0\}$, we claim that S contains some z_n; indeed, we show that S contains z_1. Now S must contain some element x of order p, and $x = z_1^c$, where $1 \leq c < p$ [for $\langle z_1 \rangle$ contains all the elements in $\mathbb{Z}(p^\infty)$ of order p]. Since p is prime, $(c, p) = 1$, and there are integers u, v with $1 = cu + pv$; hence, $z_1 = z_1^{cu+pv} = z_1^{cu} = x^u \in S$. Let d be the largest integer with $z_d \in S$. Clearly, $\langle z_d \rangle \subseteq S$. For the reverse inclusion, let $s \in S$. If s has order $p^n > p^d$, then $\langle s \rangle$ contains z_n, because $\langle z_n \rangle$ contains all the elements of order p^n in $\mathbb{Z}(p^\infty)$. But this contradicts our observation that $z_\ell \notin S$ for all $\ell > d$. Hence, s has order $\leq p^d$, and so $s \in \langle z_d \rangle$; therefore, $S = \langle z_d \rangle$.

As the only proper nonzero subgroups of $\mathbb{Z}(p^\infty)$ are the groups $\langle z_n \rangle$, it follows that the subgroups are well-ordered by inclusion.

(iv) First, $\mathbb{Z}(p^\infty)$ does not have the ACC, as the chain of subgroups

$$\{1\} \subsetneq \langle z_1 \rangle \subsetneq \langle z_2 \rangle \subsetneq \cdots$$

illustrates. It is proved in Proposition A.3 of the Appendix that every strictly decreasing sequence in a well-ordered set is finite; it follows that $\mathbb{Z}(p^\infty)$ has the DCC on subgroups. •

Notation. If G is an abelian group and n is a positive integer, then

$$G[n] = \{g \in G : ng = 0\}.$$

It is easy to see that $G[n]$ is a subgroup of G. Note that if p is prime, then $G[p]$ is a vector space over $\mathbb{F}_p$.

Lemma 9.26. *If G and H are divisible p-primary abelian groups, then $G \cong H$ if and only if $G[p] \cong H[p]$.*

Proof. If there is an isomorphism $f : G \to H$, then it is easy to see that its restriction $f|G[p]$ is an isomorphism $G[p] \to H[p]$ (whose inverse is $f^{-1}|H[p]$).

For sufficiency, assume that $f : G[p] \to H[p]$ is an isomorphism. Composing with the inclusion $H[p] \to H$, we may assume that $f : G[p] \to H$. Since H is injective, f extends to a homomorphism $F : G \to H$; we claim that any such F is an isomorphism.

(i) F is an injection.

If $g \in G$ has order p, then $F(g) = f(g) \neq 0$, by hypothesis. Suppose that g has order p^n for $n \geq 2$. If $F(g) = 0$, then $F(p^{n-1}g) = 0$, and this contradicts the hypothesis, because $p^{n-1}g$ has order p. Therefore, F is an injection.

(ii) F is a surjection.

We show, by induction on $n \geq 1$, that if $h \in H$ has order p^n, then $h \in \operatorname{im} F$. If $n = 1$, then $h \in H[p] = \operatorname{im} f \subseteq \operatorname{im} F$. For the inductive step, assume that $h \in H$ has order p^{n+1}. Now $p^n h \in H[p]$, so there exists $g \in G$ with $F(g) = f(g) = p^n h$. Since G is divisible, there is $g' \in G$ with $p^n g' = g$; thus, $p^n(h - F(g')) = 0$. By induction, there is $x \in G$ with $F(x) = h - F(g')$. Therefore, $F(x + g') = h$, as desired. •

The next theorem classifies all divisible abelian groups.

Definition. If D is a divisible abelian group, define

$$\delta_\infty(D) = \dim_{\mathbb{Q}}(D/tD)$$

and, for all primes p,

$$\delta_p(D) = \dim_{\mathbb{F}_p}(D[p]).$$

Theorem 9.27.

(i) *An abelian group D is an injective $\mathbb{Z}$-module if and only it is a divisible group.*

(ii) *Every divisible abelian group is isomorphic to a direct sum of copies of $\mathbb{Q}$ and of copies of $\mathbb{Z}(p^\infty)$ for various primes p.*

(iii) *Two divisible groups D and D' are isomorphic if and only if $\delta_\infty(D) = \delta_\infty(D')$ and $\delta_p(D) = \delta_p(D')$ for all primes p.*

Proof. (i) This is proved in Corollary 7.73.

(ii) If $x \in D$ has finite order, if n is a positive integer, and if $x = ny$, then y has finite order. It follows that if D is divisible, then its torsion subgroup tD is also divisible, and hence

$$D = tD \oplus V,$$

where V is torsion-free (by Proposition 7.64 on page 481). Since every quotient of a divisible group is divisible, V is torsion-free and divisible, and hence it is a vector space over $\mathbb{Q}$, by Proposition 9.23.

Now tD is the direct sum of its primary components: $tD = \sum_p T_p$, each of which is p-primary and divisible, and so it suffices to prove that each T_p is a direct sum of copies of $\mathbb{Z}(p^\infty)$. If $\dim(T_p[p]) = r$ (r may be infinite), define W to be a direct sum of r copies of $\mathbb{Z}(p^\infty)$, so that $\dim(W[p]) = r$. Lemma 9.26 now shows that $T_p \cong W$.

(iii) By Proposition 9.2(ii), if $D \cong D'$, then $D/tD \cong D'/tD'$ and $tD \cong tD'$; hence, the p-primary components $(tD)_p \cong (tD')_p$ for all p. But D/tD and D'/tD' are isomorphic vector spaces over $\mathbb{Q}$, and hence have the same dimension; moreover, the vector spaces $(tD)_p[p]$ and $(tD')_p[p]$ are also isomorphic, so they, too, have the same dimension.

For the converse, write $D = V \oplus \sum_p T_p$ and $D' = V' \oplus \sum_p T'_p$, where V and V' are torsion-free divisible, and T_p and T'_p are p-primary divisible. By Lemma 9.26, $\delta_p(D) = \delta_p(D')$ implies $T_p \cong T'_p$, while $\delta_\infty(D) = \delta_\infty(D')$ implies that the vector spaces V and V' are isomorphic. By Proposition 7.30, these isomorphisms can be assembled to give an isomorphism between D and D'. •

We can now describe some familiar groups, but the reader may have to review a bit of field theory.

Corollary 9.28. *Let k be an algebraically closed field, let $k^\times$ be its multiplicative group, and let T be the torsion subgroup of $k^\times$.*

(i) *If k has characteristic 0, then $T \cong \mathbb{Q}/\mathbb{Z}$, and $k^\times \cong (\mathbb{Q}/\mathbb{Z}) \oplus V$, where V is a vector space over $\mathbb{Q}$.*

(ii) *If k has prime characteristic p, then $T \cong \sum_{q \neq p} \mathbb{Z}(q^\infty)$. If k is the algebraic closure of $\mathbb{F}_p$, then*[5]

$$k^\times \cong \sum_{q \neq p} \mathbb{Z}(q^\infty).$$

Proof. Since k is algebraically closed, the polynomials $x^n - a$ have roots in k whenever $a \in k$; this says that every a has an nth root in k, which is the multiplicative way of saying that $k^\times$ is a divisible group. An element $a \in k$ has finite order if and only if $a^n = 1$ for some positive integer n; that is, a is an nth root of unity. It is easy to see that T is, itself, divisible. Hence, $k^\times = T \oplus V$, by Lemma 9.24, where V is a vector space over $\mathbb{Q}$ (for V is torsion-free divisible).

(i) If $k = \overline{\mathbb{Q}}$ is the algebraic closure of $\mathbb{Q}$, there is no loss in generality in assuming that $k \subseteq \mathbb{C}$. Now the torsion subgroup T of k consists of all the roots of unity $e^{2\pi ir}$, where $r \in \mathbb{Q}$. It follows easily that the map $r \mapsto e^{2\pi ir}$ is a surjection $\mathbb{Q} \to T$ having kernel $\mathbb{Z}$, so that $T \cong \mathbb{Q}/\mathbb{Z}$.

If k is any algebraically closed field of characteristic 0, then $\mathbb{Q} \subseteq k$ implies $\overline{\mathbb{Q}} \subseteq k$. There cannot be any roots of unity in k not in $\overline{\mathbb{Q}}$, because $\overline{\mathbb{Q}}$ already contains n roots of $x^n - 1$.

(ii) Let $k = \overline{\mathbb{F}}_p$. Every element $a \in k$ is algebraic over $\mathbb{F}_p$, and so $\mathbb{F}_p(a)/\mathbb{F}_p$ is a finite field extension; say, $[\mathbb{F}_p(a) : \mathbb{F}_p] = m$ for some m. Hence, $|\mathbb{F}_p(a)| = p^m$ and $\mathbb{F}_p(a)$ is a finite field. Now every nonzero element in a finite field is a root of unity (for it is a root of $x^{p^m} - x$ for some m). But $k^\times = T \oplus V$, where V is a vector space over $\mathbb{Q}$. It follows that $V = \{0\}$, for every nonzero element of k is a root of unity.

We now examine the primary components of $k^\times$. If $q \neq p$ is a prime, then the polynomial $f(x) = x^q - 1$ has no repeated roots (for $\gcd(f(x), f'(x)) = 1$), and so there is some qth root of unity other than 1. Thus, the q-primary component of $k^\times$ is nontrivial, and so there is at least one summand isomorphic to $\mathbb{Z}(q^\infty)$. Were there more than one such summand, there would be more than q elements of order q, and this would provide too many roots for $x^q - 1$ in the field k. Finally, there is no summand isomorphic to $\mathbb{Z}(p^\infty)$, for the polynomial $x^p - 1 = (x-1)^p$ in $k[x]$, and so it has no roots other than 1. •

Corollary 9.29. *The following abelian groups are isomorphic:*

$$\mathbb{C}^\times; \quad (\mathbb{Q}/\mathbb{Z}) \oplus \mathbb{R}; \quad \mathbb{R}/\mathbb{Z}; \quad \prod_p \mathbb{Z}(p^\infty); \quad S^1.$$

[5]The additive group of k is easy to describe, for k is a vector space over $\mathbb{F}_p$, and so it is a direct sum of (infinitely many) copies of $\mathbb{F}_p$.

Here S^1 is the circle group; that is, the multiplicative group of all complex numbers z with $|z| = 1$.

Proof. The reader may use Theorem 9.27, because, for every group G on the list, we have $\delta_p(G) = 1$ for all primes p and $\delta_\infty(G) = c$ (the cardinal of the continuum). See Exercise 9.29 on page 666 for $G = \prod_p \mathbb{Z}(p^\infty)$. •

EXERCISES

9.1 Let $G = \prod_p \langle a_p \rangle$, where p varies over all the primes, and $\langle a_p \rangle \cong \mathbb{I}_p$.

(i) Prove that $tG = \sum_p \langle a_p \rangle$.
Hint. Use Exercise 5.4 on page 267.

(ii) Prove that G/tG is a divisible group.

(iii) Prove that tG is not a direct summand of G.
Hint. Show that $\operatorname{Hom}(\mathbb{Q}, G) = \{0\}$ but that $\operatorname{Hom}(\mathbb{Q}, G/tG) \neq \{0\}$, and conclude that G/tG cannot be isomorphic to a subgroup of G.

9.2 Let R be a PID, and let M be an R-module, not necessarily primary. Define a submodule $S \subseteq M$ to be a ***pure submodule*** if $S \cap rM = rS$ for all $r \in R$.

(i) Prove that if M is a (p)-primary module, where (p) is a nonzero prime ideal in R, then a submodule $S \subseteq M$ is pure as just defined if and only if $S \cap p^n M = p^n S$ for all $n \geq 0$.

(ii) Prove that every direct summand of M is a pure submodule.

(iii) Prove that the torsion submodule tM is a pure submodule of M.

(iv) Prove that if M/S is torsion-free, then S is a pure submodule of M.

(v) Prove that if $\mathcal{S}$ is a family of pure submodules of a module M that is a chain under inclusion (that is, if $S, S' \in \mathcal{S}$, then either $S \subseteq S'$ or $S' \subseteq S$), then $\bigcup_{S \in \mathcal{S}} S$ is a pure submodule of M.

(vi) Give an example of a pure submodule that is not a direct summand.

9.3 (i) If F is a finitely generated free R-module, where R is a PID, prove that every pure submodule of F is a direct summand.

(ii) If R is a PID and M is a finitely generated R-module, prove that a submodule $S \subseteq M$ is a pure submodule of M if and only if S is a direct summand of M.

9.4 Prove that if R is a domain that is not a field, then an R-module M that is both projective and injective must be $\{0\}$.
Hint. Use Exercise 7.43 on page 487.

9.5 If M is a torsion module over a domain R, prove that

$$\operatorname{Hom}_R(M, M) \cong \prod_P \operatorname{Hom}_R(M_P, M_P),$$

where M_P is the P-primary component of M.

9.6 (i) If G is a torsion group with p-primary components $\{G_p : p \in P\}$, where P is the set of all primes, prove that $G = t\big(\prod_{p \in P} G_p\big)$.

(ii) Prove that $\big(\prod_{p \in P} G_p\big)/\big(\sum_{p \in P} G_p\big)$ is torsion-free and divisible.
Hint. Use Exercise 5.4 on page 267.

9.7 If M is an R-module, where R is a domain, and if $r \in R$, let $\mu_r \colon M \to M$ be multiplication by r; that is, $\mu_r \colon m \mapsto rm$ [see Example 7.2(iii)].

(i) If $Q = \operatorname{Frac}(R)$, prove that an R-module is a vector space over Q if and only if M is torsion-free and divisible.

(ii) Prove that μ_r is an injection for every $r \neq 0$ if and only if M is torsion-free.

(iii) Prove that μ_r is a surjection for every $r \neq 0$ if and only if M is divisible.

(iv) Prove that M is a vector space over Q if and only if, for every $r \neq 0$, the map $\mu_r \colon M \to M$ is an isomorphism.

9.8 (i) Let R be a domain, let $r \in R$, and let M be an R-module. If $\mu_r \colon M \to M$ is multiplication by r, prove, for every R-module A, that the induced maps

$$(\mu_r)_* \colon \operatorname{Hom}_R(A, M) \to \operatorname{Hom}_R(A, M)$$

and

$$(\mu_r)^* \colon \operatorname{Hom}_R(M, A) \to \operatorname{Hom}_R(M, A)$$

are also multiplication by r.

(ii) Let R be a domain with $Q = \operatorname{Frac}(R)$. Using Exercise 9.7 on page 664, prove, for every R-module M, that both $\operatorname{Hom}_R(Q, M)$ and $\operatorname{Hom}_R(M, Q)$ are vector spaces over Q.

9.9 (i) If M and N are finitely generated torsion R-modules, prove, for all primes P and all $n \geq 0$, that

$$U_P(n, M \oplus N) = U_P(n, M) + U_P(n, N),$$

(ii) If A, B, and C are finitely generated R-modules, where R is a PID, prove that $A \oplus B \cong A \oplus C$ implies $B \cong C$.

(iii) If A and B are finitely generated R-modules, where R is a PID, prove that $A \oplus A \cong B \oplus B$ implies $A \cong B$.

9.10 If A is an abelian group, call a subset X of A ***linearly independent*** if, whenever $\sum_i m_i x_i = 0$, where $m_i \in \mathbb{Z}$ and almost all $m_i = 0$, then $m_i = 0$ for all i. Define $\operatorname{rank}(A)$ to be the number of elements in a maximal linearly independent subset of A.

(i) If X is linearly independent, prove that $\langle X \rangle = \sum_{x \in X} \langle x \rangle$, a direct sum of cyclic groups.

(ii) If A is torsion, prove that $\operatorname{rank}(A) = 0$.

(iii) If A is free abelian, prove that the two notions of rank coincide [the earlier notion defined $\operatorname{rank}(A)$ as the number of elements in a basis of A].

(iv) Prove that $\operatorname{rank}(A) = \dim(\mathbb{Q} \otimes_{\mathbb{Z}} A)$, and conclude that every two maximal linearly independent subsets of A have the same number of elements; that is, $\operatorname{rank}(A)$ is well-defined.

(v) If $0 \to A \to B \to C \to 0$ is an exact sequence of abelian groups, prove that $\operatorname{rank}(B) = \operatorname{rank}(A) + \operatorname{rank}(C)$.

9.11 (***Kulikov***) If G is an abelian p-group, call a subset $X \subseteq G$ ***pure-independent*** if X is linearly independent (see Exercise 9.10) and $\langle X \rangle$ is a pure subgroup.

(i) Prove that G has a maximal pure-independent subset.

(ii) If X is a maximal pure-independent subset of G, the subgroup $B = \langle X \rangle$ is called a ***basic subgroup*** of G. Prove that if B is a basic subgroup of G, then G/B is divisible.

9.12 Prove that if G and H are torsion abelian groups, then $G \otimes_{\mathbb{Z}} H$ is a direct sum of cyclic groups.
Hint. Use an exact sequence $0 \to B \to G \to G/B \to 0$, where B is a basic subgroup, along with the following theorem proved in Rotman, *An Introduction to Homological Algebra*,

pages 94–96): If $0 \to A' \xrightarrow{i} A \to A'' \to 0$ is an exact sequence of abelian groups and if $i(A')$ is a pure subgroup of A, then, for every abelian group E, there is exactness of

$$0 \to A' \otimes_{\mathbb{Z}} E \to A \otimes_{\mathbb{Z}} E \to A'' \otimes_{\mathbb{Z}} E \to 0.$$

9.13 Let M be a P-primary R-module, where R is a PID and $P = (p)$ is a prime ideal. Define, for all $n \geq 0$,

$$V_P(n, M) = \dim\Big((p^n M \cap M[p])/(p^{n+1} M \cap M[p])\Big),$$

where $M[p] = \{m \in M : pm = 0\}$. (This invariant is introduced because we cannot subtract infinite cardinal numbers.)

(i) Prove that $V_P(n, M) = U_P(n, M)$ when M is finitely generated

(ii) Let $M = \sum_{i \in I} C_i$ be a direct sum of cyclic modules C_i, where I is any index set, possibly infinite. Prove that the number of summands C_i having order ideal (p^n) is $V_P(n, M)$, and hence it is an invariant of M.

(iii) Let M and M' be torsion modules that are direct sums of cyclic modules. Prove that $M \cong M'$ if and only if $V_P(n, M) = V_P(n, M')$ for all $n \geq 0$ and all prime ideals P.

9.14 (i) If p is a prime and $G = t\big(\prod_{k \geq 1} \langle a_k \rangle\big)$, where $\langle a_k \rangle$ is a cyclic group of order p^k, prove that G is an uncountable p-primary abelian group with $V_p(n, G) = 1$ for all $n \geq 0$.

(ii) Use Exercise 9.13 to prove that the primary group G in part (i) is not a direct sum of cyclic groups.

9.15 Generalize Proposition 8.95 as follows: If R is a domain, D is a divisible R-module, and T is a torsion R-module with every element of finite order, then $D \otimes_R T = \{0\}$.

9.16 Prove that there is an additive functor $d \colon \mathbf{Ab} \to \mathbf{Ab}$ that assigns to each group G its maximal divisible subgroup dG.

9.17 (i) Prove that $\mathbb{Z}(p^\infty)$ has no maximal subgroups.

(ii) Prove that $\mathbb{Z}(p^\infty) \cong \varinjlim \mathbb{I}_{p^n}$.

(iii) Prove that a presentation of $\mathbb{Z}(p^\infty)$ is

$$(a_n,\ n \geq 1 \mid pa_1 = 0,\ pa_{n+1} = a_n \text{ for } n \geq 1).$$

9.18 Prove that an abelian group G is reduced if and only if $\mathrm{Hom}_{\mathbb{Z}}(\mathbb{Q}, G) = \{0\}$.

9.19 If $0 \to A \to B \to C \to 0$ is exact and both A and C are reduced, prove that B is reduced.
Hint. Use left exactness of $\mathrm{Hom}_{\mathbb{Z}}(\mathbb{Q},\ )$.

9.20 If $\{D_i : i \in I\}$ is a family of divisible abelian groups, prove that $\prod_{i \in I} D_i$ is isomorphic to a direct *sum* of divisible groups.

9.21 Prove that $\mathbb{Q}^\times \cong \mathbb{I}_2 \oplus F$, where F is a free abelian group of infinite rank.

9.22 Prove that $\mathbb{R}^\times \cong \mathbb{I}_2 \oplus \mathbb{R}$.
Hint. Use e^x.

9.23 (i) Prove, for every group homomorphism $f \colon \mathbb{Q} \to \mathbb{Q}$, that there exists $r \in \mathbb{Q}$ with $f(x) = rx$ for all $x \in \mathbb{Q}$.

(ii) Prove that $\mathrm{Hom}_{\mathbb{Z}}(\mathbb{Q}, \mathbb{Q}) \cong \mathbb{Q}$.

(iii) Prove that $\mathrm{End}(\mathbb{Q}) \cong \mathbb{Q}$ as rings.

9.24 For every abelian group A, prove that $\mathrm{Hom}_{\mathbb{Z}}(A, \mathbb{Q})$ and $\mathrm{Hom}_{\mathbb{Z}}(\mathbb{Q}, A)$ are vector spaces over $\mathbb{Q}$.

9.25 Prove that if G is a nonzero abelian group, then $\text{Hom}_{\mathbb{Z}}(G, \mathbb{Q}/\mathbb{Z}) \neq \{0\}$.

9.26 Prove that an abelian group G is injective if and only if every nonzero quotient group is infinite.

9.27 Prove that if G is an infinite abelian group all of whose proper subgroups are finite, then $G \cong \mathbb{Z}(p^\infty)$ for some prime p.[6]

9.28 (i) Let $D = \sum_{i=1}^{n} D_i$, where each $D_i \cong \mathbb{Z}(p_i^\infty)$ for some prime p_i. Prove that every subgroup of D has DCC.

(ii) Prove, conversely, that if an abelian group G has DCC, then G is isomorphic to a subgroup of a direct sum of a finite number of copies of $\mathbb{Z}(p_i^\infty)$.

9.29 Let $G = \prod_{p \in P} \mathbb{Z}(p^\infty)$, where P is the set of all primes. Prove that $\delta_p(G) = 1$ for all $p \in P$, and that $\delta_\infty(G) = c$, where c is the cardinal of the continuum.

Hint. Use Exercise 9.6 on page 663 after noting that $\prod_{p \in P} \mathbb{Z}(p^\infty)$ has cardinality c while $\sum_{p \in P} \mathbb{Z}(p^\infty)$ is countable.

9.30 Let $R = k[x, y]$ be the polynomial ring in two variables over a field k, and let $I = (x, y)$.

(i) Prove that $x \otimes y - y \otimes x \neq 0$ in $I \otimes_R I$.

Hint. Show that this element has a nonzero image in $(I/I^2) \otimes_R (I/I^2)$.

(ii) Prove that $x \otimes y - y \otimes x$ is a torsion element in $I \otimes_R I$, and conclude that the tensor product of torsion-free modules need not be torsion-free.

9.31 Let $\mathcal{C}$ be the category of all finitely generated R-modules, where R is a PID.

(i) Compute the Grothendieck group $K_0(\mathcal{C})$.

(ii) Compute the Grothendieck group $K'(\mathcal{C})$.

9.2 Rational Canonical Forms

In Chapter 3, we saw that if $T : V \to V$ is a linear transformation and if $X = x_1, \ldots, x_n$ is a basis of V, then T determines the matrix $A = {}_X[T]_X$ whose ith column consists of the coordinate-set of $T(x_i)$ with respect to X. If Y is another basis of V, then the matrix $B = {}_Y[T]_Y$ may be different from A. On the other hand, Corollary 3.101 on page 176 says that two matrices A and B arise from the same linear transformation if and only if A and B are *similar*; that is, there exists a nonsingular matrix P with $B = PAP^{-1}$.

Corollary 3.101. *Let $T : V \to V$ be a linear transformation on a vector space V over a field k. If X and Y are bases of V, then there is a nonsingular matrix P with entries in k so that*

$$ {}_Y[T]_Y = P\big({}_X[T]_X\big)P^{-1}. $$

Conversely, if $B = PAP^{-1}$, where B, A, and P are $n \times n$ matrices with entries in k and P is nonsingular, then there is a linear transformation $T : k^n \to k^n$ and bases X and Y of k^n such that $B = {}_Y[T]_Y$ and $A = {}_X[T]_X$.

We now consider how to determine whether two given matrices are similar; that is, whether they arise from the same linear transformation.

[6] There exist infinite nonabelian groups all of whose proper subgroups are finite. Indeed, ***Tarski monsters*** exist: These are infinite groups all of whose proper subgroups have prime order.

Example 9.30.
Recall Example 7.1(v) on page 424: If $T\colon V \to V$ is a linear transformation, where V is a vector space over a field k, then V admits a scalar multiplication by polynomials $f(x) \in k[x]$:

$$f(x)v = \Big(\sum_{i=0}^{m} c_i x^i\Big)v = \sum_{i=0}^{m} c_i T^i(v),$$

where T^0 is the identity map 1_V, and T^i is the composite of T with itself i times if $i \geq 1$. We denote this $k[x]$-module by V^T.

We now show that if V is n-dimensional, then the $k[x]$-module V^T is a torsion module. By Corollary 3.88, for each $v \in V$, the list $v, T(v), T^2(v), \ldots, T^n(v)$ must be linearly dependent (for it contains $n+1$ vectors). Therefore, there are $c_i \in k$, not all 0, with $\sum_{i=0}^{n} c_i T^i(v) = 0$; but this says that $g(x) = \sum_{i=0}^{n} c_i x^i$ lies in the order ideal $\operatorname{ann}(v)$. ◂

There is an important special case of the construction of the $k[x]$-module V^T. If A is an $n \times n$ matrix with entries in k, define $T : k^n \to k^n$ by $T(v) = Av$ (recall that the elements of k^n are $n \times 1$ column vectors v, so that Av is matrix multiplication). We denote the $k[x]$-module $(k^n)^T$ by $(k^n)^A$; thus, the action is given by

$$f(x)v = \Big(\sum_{i=0}^{m} c_i x^i\Big)v = \sum_{i=0}^{m} c_i A^i v.$$

We now interpret the results in the previous section about modules over general PIDs for the $k[x]$-modules V^T and $(k^n)^A$. If $T\colon V \to V$ is a linear transformation, then a submodule W of V^T is an ***invariant subspace***; that is, W is a subspace of V with $T(W) \subseteq W$, and so the restriction $T|W$ is a linear transformation on W; that is, $T|W\colon W \to W$.

Definition. If A is an $r \times r$ matrix and B is an $s \times s$ matrix, then their ***direct sum*** $A \oplus B$ is the $(r+s) \times (r+s)$ matrix

$$A \oplus B = \begin{bmatrix} A & 0 \\ 0 & B \end{bmatrix}.$$

Lemma 9.31. *If $V^T = W \oplus W'$, where W and W' are submodules, then*

$${}_{B\cup B'}[T]_{B\cup B'} = {}_B[T|W]_B \oplus {}_{B'}[T|W']_{B'},$$

where $B = w_1, \ldots, w_r$ is a basis of W and $B' = w_1', \ldots, w_s'$ is a basis of W'.

Proof. Since W and W' are submodules, we have $T(W) \subseteq W$ and $T(W') \subseteq W'$; that is, the restrictions $T|W$ and $T|W'$ are linear transformations on W and W', respectively. Since $V = W \oplus W'$, the union $B \cup B'$ is a basis of V. Finally, ${}_{B\cup B'}[T]_{B\cup B'}$ is a direct sum as in the statement of the lemma: $T(w_i) \in W$, so that it is a linear combination of $w_1, \ldots, w_r$, and hence it requires no nonzero coordinates from the w_j'; similarly, $T(w_j') \in W'$, and so its coordinates from the w_i are all 0. •

When we studied permutations, we saw that the cycle notation allowed us to recognize important properties that are masked by the conventional functional notation. We now ask whether there is an analogous way to denote matrices; more precisely, if V^T is a cyclic $k[x]$-module, can we find a basis B of V so that the corresponding matrix ${}_B[T]_B$ displays important properties of T?

Lemma 9.32. *A submodule W of V^T is cyclic of finite order if and only if there is a vector $v \in W$ and an integer $s \geq 1$ such that*

$$v, Tv, T^2v, \ldots, T^{s-1}v$$

is a basis of W. Moreover, if

$$T^s v + \sum_{i=0}^{s-1} c_i T^i v = 0,$$

then the order ideal $\operatorname{ann}(v) = (g)$*, where* $g(x) = x^s + c_{s-1}x^{s-1} + \cdots + c_1 x + c_0$; *that is,*

$$W \cong k[x]/(g).$$

Proof. Assume that $W = \langle v \rangle = \{f(x)v : f(x) \in k[x]\}$. Since V, hence W, is finite-dimensional, there is an integer $s \geq 1$ and a linearly independent list $v, Tv, T^2v, \ldots, T^{s-1}v$ that becomes linearly dependent when we adjoin $T^s v$. Hence, there are $c_i \in k$ with

$$T^s v + \sum_{i=0}^{s-1} c_i T^i v = 0.$$

If $w \in W$, then $w = f(x)v$ for some $f(x) \in k[x]$. An easy induction on $\deg(f)$ shows that w lies in the subspace spanned by $v, Tv, T^2v, \ldots, T^{s-1}v$; it follows that this list is a basis of W.

To prove the converse, assume that there is a vector $v \in W$ and an integer $s \geq 1$ such that the list $v, Tv, T^2v, \ldots, T^{s-1}v$ is a basis of W. Clearly, $W \subseteq \langle v \rangle$, the cyclic submodule generated by v. The reverse inclusion is obvious, for we are assuming that W is a submodule; hence, $f(x)v \in W$ for every $f(x) \in k[x]$.

The polynomial $g(x)$ lies in the order ideal $\operatorname{ann}(v)$. If $h(x) \in \operatorname{ann}(v)$, the division algorithm gives $q(x)$ and $r(x)$ with $h = gq + r$, where $r = 0$ or $\deg(r) < \deg(g) = s$. But $r(x) \in \operatorname{ann}(v)$, so that $r(x) = \sum_{j=0}^{t} c_j x^j$. Hence, $\sum_{j=0}^{t} c_j T^j v = 0$, where $t \leq s-1$, and this contradicts the linear independence of the basis. Therefore, $g(x)$ has smallest degree of all polynomials in $\operatorname{ann}(v)$, so that $\operatorname{ann}(v) = (g)$. Therefore, $W \cong k[x]/\operatorname{ann}(v) = k[x]/(g)$. •

Definition. If $g(x) = x + c_0$, then its ***companion matrix*** $C(g)$ is the 1×1 matrix $[-c_0]$; if $s \geq 2$ and $g(x) = x^s + c_{s-1}x^{s-1} + \cdots + c_1 x + c_0$, then its ***companion matrix*** $C(g)$ is

the $s \times s$ matrix

$$\begin{bmatrix} 0 & 0 & 0 & \cdots & 0 & -c_0 \\ 1 & 0 & 0 & \cdots & 0 & -c_1 \\ 0 & 1 & 0 & \cdots & 0 & -c_2 \\ 0 & 0 & 1 & \cdots & 0 & -c_3 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & -c_{s-1} \end{bmatrix}.$$

Obviously, we can recapture the polynomial $g(x)$ from the last column of the companion matrix $C(g)$.

Lemma 9.33. *Let $T : V \to V$ be a linear transformation on a vector space V over a field k, and let V^T be a cyclic $k[x]$-module with generator v. If the order ideal* $\operatorname{ann}(v) = (g)$, *where $g(x) = x^s + c_{s-1}x^{s-1} + \cdots + c_1 x + c_0$, then $B = v, Tv, T^2v, \ldots, T^{s-1}v$ is a basis of V and the matrix ${}_B[T]_B$ is the companion matrix $C(g)$.*

Proof. Let $A = {}_B[T]_B$. By definition, the first column of A consists of the coordinates of $T(v)$, the second column the coordinates of $T(Tv) = T^2v$, and, more generally, if $i < s-1$, then $T(T^i v) = T^{i+1}v$; that is, T sends each basis vector into the next one. However, on the last basis vector, $T(T^{s-1}v) = T^s v$. But $T^s v = -\sum_{i=0}^{s-1} c_i T^i v$, where $g(x) = x^s + \sum_{i=0}^{s-1} c_i x^i$. Thus, ${}_B[T]_B$ is the companion matrix $C(g)$. •

Theorem 9.34.

(i) *Let A be an $n \times n$ matrix with entries in a field k. If*

$$(k^n)^A = W_1 \oplus \cdots \oplus W_r,$$

where each W_i is cyclic, say, with order ideal (f_i), then A is similar to a direct sum of companion matrices

$$C(f_1) \oplus \cdots \oplus C(f_r).$$

(ii) *Every $n \times n$ matrix A over a field k is similar to a direct sum of companion matrices*

$$C(g_1) \oplus \cdots \oplus C(g_t)$$

in which the $g_i(x)$ are monic polynomials and

$$g_1(x) \mid g_2(x) \mid \cdots \mid g_t(x).$$

Proof. Define $V = k^n$ and define $T : V \to V$ by $T(y) = Ay$, where y is a column vector. (i) By Lemma 9.33, each W_i has a basis $B_i = v_i, Tv_i, T^2v_i, \ldots$ and, with respect to this basis B_i, the restriction $T|W_i$ has matrix $C(f_i)$, the companion matrix of $f_i(x)$. With respect to the basis $B_1 \cup \cdots \cup B_r$, the transformation T has the desired matrix, by Proposition 9.31. Finally, A is similar to $C(f_1) \oplus \cdots \oplus C(f_r)$, by Corollary 3.101.

(ii) By Example 9.30, the finitely generated $k[x]$-module V^T is a torsion module, and so the consequence of the basis theorem, Proposition 9.18, gives

$$(k^n)^A = W_1 \oplus W_2 \oplus \cdots \oplus W_t,$$

where each W_i is cyclic, say, with generator v_i having order ideal (g_i), and $g_1(x) \mid g_2(x) \mid \cdots \mid g_t(x)$. The statement now follows from part (i). •

Definition. A ***rational***[7] ***canonical***[8] ***form*** is a matrix R that is a direct sum of companion matrices,

$$R = C(g_1) \oplus \cdots \oplus C(g_t),$$

where the $g_i(x)$ are monic polynomials with $g_1(x) \mid g_2(x) \mid \cdots \mid g_t(x)$.

If a matrix A is similar to the rational canonical form

$$C(g_1) \oplus \cdots \oplus C(g_t),$$

where $g_1(x) \mid g_2(x) \mid \cdots \mid g_t(x)$, then we say that the ***invariant factors*** of A are $g_1(x), g_2(x), \ldots, g_t(x)$.

We have just proved that every $n \times n$ matrix over a field is similar to a rational canonical form, and so it has invariant factors. Can a matrix A have more than one list of invariant factors?

Theorem 9.35. *Two $n \times n$ matrices A and B with entries in a field k are similar if and only if they have the same invariant factors. Moreover, a matrix is similar to exactly one rational canonical form.*

Proof. By Corollary 3.101, A and B are similar if and only if $(k^n)^A \cong (k^n)^B$. By Theorem 9.22, $(k^n)^A \cong (k^n)^B$ if and only if their invariant factors are the same.

There is only one rational canonical form for a given list of invariant factors $g_1(x)$, $g_2(x), \ldots, g_t(x)$, namely, $C(g_1) \oplus \cdots \oplus C(g_t)$. If a matrix were similar to two distinct rational canonical forms, then it would have two different lists of invariant factors, contrary to the first statement of this theorem. •

Here is a theorem analogous to Corollary 3.41, which states that if k is a subfield of a field K and if $f(x), g(x) \in k[x]$, then their gcd in $k[x]$ is equal to their gcd in $K[x]$.

---

[7] If $E \subseteq \mathbb{R}$ is an extension of $\mathbb{Q}$, then every element $e \in E$ that is not in $\mathbb{Q}$ is irrational. More generally, if E/k is a field extension, then we call the elements of the ground field k *rational*. This is the usage of the adjective *rational* in *rational canonical form*, for all the entries of a rational canonical form lie in the field k and not in some extension of it. In contrast, the Jordan canonical form, to be discussed in the next section, involves the eigenvalues of a matrix that may not lie in k.

[8] The adjective *canonical* originally meant something dictated by ecclesiastical law, as *canonical hours* being those times devoted to prayers. The meaning broadened to mean things of excellence, leading to the mathematical meaning of something given by a general rule or formula.

Corollary 9.36.

(i) *Let k be a subfield of a field K, and let A and B be $n \times n$ matrices with entries in k. If A and B are similar over K, then they are similar over k (i.e., if there is a nonsingular matrix P having entries in K with $B = PAP^{-1}$, then there is a nonsingular matrix Q having entries in k with $B = QAQ^{-1}$).*

(ii) *If $\overline{k}$ is the algebraic closure of a field k, then two $n \times n$ matrices A and B with entries in k are similar over k if and only if they are similar over $\overline{k}$.*

Proof. (i) Suppose that $g_1(x), \ldots, g_t(x)$ are the invariant factors of A regarded as a matrix over k, while $G_1(x), \ldots, G_q(x)$ are the invariant factors of A regarded as a matrix over K. By the theorem, the two lists of polynomials coincide, for both are invariant factors for A as a matrix over K.

Now B has the same invariant factors as A, for they are similar over K; since these invariant factors lie in k, however, A and B are similar over k.

(ii) Immediate from part (i). •

For example, suppose that A and B are matrices with real entries that are similar over the complexes; that is, if there is a nonsingular complex matrix P such that $B = PAP^{-1}$, then there exists a nonsingular real matrix Q such that $B = QAQ^{-1}$.

The first step in analyzing a matrix A is to see whether it leaves any one-dimensional subspaces of k^n invariant; that is, are there any nonzero vectors x with $Ax = \alpha x$ for some scalar α? We call α an ***eigenvalue*** of A and we call x an ***eigenvector*** of A for α. To say that $Ax = \alpha x$ for x nonzero is to say that x is a nontrivial solution of the homogeneous system $(A - \alpha I)x = 0$; that is, $A - \alpha I$ is a singular matrix. But a matrix with entries in a field is singular if and only if its determinant is 0. Recall that the ***characteristic polynomial*** of A is $\psi_A(x) = \det(xI - A) \in k[x]$,[9] and so the eigenvalues of A are the roots of $\psi_A(x)$. If $\overline{k}$ is the algebraic closure of k, then $\psi_A(x) = \prod_{i=1}^{n}(x - \alpha_i)$, and so the constant term of $\psi_A(x)$ is $(-1)^n \prod \alpha_i$. On the other hand, the constant term of any polynomial $f(x)$ is just $f(0)$; setting $x = 0$ in $\psi_A(x) = \det(xI - A)$ gives $\psi_A(0) = (-1)^n \det(A)$. It follows that $\det(A)$ is the product of the eigenvalues.

Here are some elementary facts about eigenvalues.

Corollary 9.37. *Let A be an $n \times n$ matrix with entries in a field k.*

(i) *A is singular if and only if 0 is an eigenvalue of A.*

(ii) *If α is an eigenvalue of A, then α^n is an eigenvalue of A^n.*

(iii) *If A is nonsingular and α is an eigenvalue of A, then $\alpha \neq 0$ and α^{-1} is an eigenvalue of A^{-1}.*

[9] We continue using familiar properties of determinants even though they will not be proved until Section 9.9.

Proof. (i) If A is singular, then the homogeneous system $Ax = 0$ has a nontrivial solution; that is, there is a nonzero x with $Ax = 0$. But this just says that $Ax = 0x$, and so 0 is an eigenvalue.

Conversely, if 0 is an eigenvalue, then $0 = \det(0I - A) = \pm\det(A)$, so that $\det(A) = 0$ and A is singular.

(ii) There is a nonzero vector v with $Av = \alpha v$. It follows by induction on $n \geq 1$ that $A^n v = \alpha^n v$.

(iii) If x is an eigenvector for A and α, then

$$x = A^{-1}Ax = A^{-1}\alpha x = \alpha A^{-1}x.$$

Therefore, $\alpha \neq 0$ (because eigenvectors are nonzero) and $\alpha^{-1}x = A^{-1}x$. •

Let us return to canonical forms.

Lemma 9.38. *If $g(x) \in k[x]$, then $\det\bigl(xI - C(g)\bigr) = g(x)$.*

Proof. If $\deg(g) = s \geq 2$, then

$$xI - C(g) = \begin{bmatrix} x & 0 & 0 & \cdots & 0 & c_0 \\ -1 & x & 0 & \cdots & 0 & c_1 \\ 0 & -1 & x & \cdots & 0 & c_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -1 & x + c_{s-1} \end{bmatrix},$$

and Laplace expansion across the first row gives

$$\det(xI - C(g)) = x\det(L) + (-1)^{1+s}c_0\det(M),$$

where L is the matrix obtained by erasing the top row and the first column, and M is the matrix obtained by erasing the top row and the last column. Now M is a triangular $(s-1) \times (s-1)$ matrix having -1's on the diagonal, while $L = xI - C\bigl((g(x) - c_0)/x\bigr)$. By induction, $\det(L) = (g(x) - c_0)/x$, while $\det(M) = (-1)^{s-1}$. Therefore,

$$\det(xI - C(g)) = x[(g(x) - c_0)/x] + (-1)^{(1+s)+(s-1)}c_0 = g(x). \quad \bullet$$

If $R = C(g_1) \oplus \cdots \oplus C(g_t)$ is a rational canonical form, then

$$xI - R = \bigl[xI - C(g_1)\bigr] \oplus \cdots \oplus \bigl[xI - C(g_t)\bigr],$$

and so the lemma and Proposition 9.163, which says that $\det(B_1 \oplus \cdots \oplus B_t) = \prod_i \det(B_i)$, give

$$\psi_R(x) = \prod_{i=1}^{t} \psi_{C(g_i)}(x) = \prod_{i=1}^{t} g_i(x).$$

Thus, the characteristic polynomial is the product of the invariant factors; in light of Corollary 9.20, the characteristic polynomial of an $n \times n$ matrix A over a field k is the analog for $(k^n)^A$ of the order of a finite abelian group.

Example 9.39.
We now show that similar matrices have the same characteristic polynomial. If $B = PAP^{-1}$, then since xI commutes with every matrix, we have $P(xI) = (xI)P$ and, hence, $P(xI)P^{-1} = (xI)PP^{-1} = xI$. Therefore,

$$\begin{aligned}\psi_B(x) &= \det(xI - B)\\ &= \det(PxIP^{-1} - PAP^{-1})\\ &= \det(P[xI - A]P^{-1})\\ &= \det(P)\det(xI - A)\det(P^{-1})\\ &= \det(xI - A)\\ &= \psi_A(x).\end{aligned}$$

It follows that if A is similar to $C(g_1) \oplus \cdots \oplus C(g_t)$, then

$$\psi_A(x) = \prod_{i=1}^{t} g_i(x).$$

Therefore, similar matrices have the same eigenvalues with multiplicities. ◀

Theorem 9.40 (Cayley–Hamilton). *If A is an $n \times n$ matrix with characteristic polynomial $\psi_A(x) = x^n + b_{n-1}x^{n-1} + \cdots + b_1x + b_0$, then $\psi_A(A) = 0$; that is,*

$$A^n + b_{n-1}A^{n-1} + \cdots + b_1A + b_0I = 0.$$

Proof. We may assume that $A = C(g_1) \oplus \cdots \oplus C(g_t)$ is a rational canonical form, by Example 9.39, where $\psi_A(x) = g_1(x)\cdots g_t(x)$. If we regard k^n as the $k[x]$-module $(k^n)^A$, then Corollary 9.19 says that $g_t(A)y = 0$ for all $y \in k^n$. Thus, $g_t(A) = 0$. As $g_t(x) \mid \psi_A(x)$, however, we have $\psi_A(A) = 0$. •

The Cayley–Hamilton theorem is the analog of Corollary 2.44.

Definition. The ***minimum polynomial*** $m_A(x)$ of an $n \times n$ matrix A is the monic polynomial $f(x)$ of least degree with the property that $f(A) = 0$.

Proposition 9.41. *The minimum polynomial $m_A(x)$ is a divisor of the characteristic polynomial $\psi_A(x)$, and every eigenvalue of A is a root of $m_A(x)$.*

Proof. The Cayley–Hamilton theorem shows that $m_A(x) \mid \psi_A(x)$, while Corollary 9.19 implies that $c_t(x)$ is the minimum polynomial of A, where $c_t(x)$ is the invariant factor of A of highest degree. It follows from the fact that

$$\psi_A(x) = c_1(x)\cdots c_t(x),$$

where $c_1(x) \mid c_2(x) \mid \cdots \mid c_t(x)$, that $m_A(x) = c_t(x)$ is a polynomial having every eigenvalue of A as a root [of course, the multiplicity as a root of $m_A(x)$ may be less than its multiplicity as a root of the characteristic polynomial $\psi_A(x)$]. •

Corollary 9.42. *If all the eigenvalues of an $n \times n$ matrix A are distinct, then $m_A(x) = \psi_A(A)$; that is, the minimum polynomial coincides with the characteristic polyomial.*

Proof. This is true because every root of $\psi_A(x)$ is a root of $m_A(x)$. •

Corollary 9.43.

(i) *An $n \times n$ matrix A is similar to a companion matrix if and only if*

$$m_A(x) = \psi_A(x).$$

(ii) *A finite abelian group G is cyclic if and only if its exponent equals its order.*

Proof. (i) A companion matrix $C(g)$ has only one invariant factor, namely, $g(x)$; but Corollary 9.19 identifies the minimum polynomial as the last invariant factor.

If $m_A(x) = \psi_A(x)$, then A has only one invariant factor, namely, $\psi_A(x)$, by Corollary 9.20. Hence, A and $C(\psi_A(x))$ have the same invariant factors, and so they are similar.

(ii) A cyclic group of order n has only one invariant factor, namely, n; but Corollary 9.19 identifies the exponent as the last invariant factor.

If the exponent of G is equal to its order $|G|$, then G has only one invariant factor, namely, $|G|$. Hence, G and $\mathbb{I}_{|G|}$ have the same invariant factors, and so they are isomorphic. •

EXERCISES

9.32 (i) How many 10×10 matrices A over $\mathbb{R}$ are there, to similarity, with $A^2 = I$?

(ii) How many 10×10 matrices A over $\mathbb{F}_p$ are there, to similarity, with $A^2 = I$?

Hint. The answer depends on whether p is odd or $p = 2$.

9.33 Find the rational canonical forms of

$$A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}, \qquad B = \begin{bmatrix} 2 & 0 & 0 \\ 1 & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix}, \qquad \text{and} \qquad C = \begin{bmatrix} 2 & 0 & 0 \\ 1 & 2 & 0 \\ 0 & 1 & 2 \end{bmatrix}.$$

9.34 If A is similar to A' and B is similar to B', prove that $A \oplus B$ is similar to $A' \oplus B'$.

9.35 Let k be a field, and let $f(x)$ and $g(x)$ lie in $k[x]$. If $g(x) \mid f(x)$ and if every root of $f(x)$ is a root of $g(x)$, show that there exists a matrix A having minimum polynomial $m_A(x) = g(x)$ and characteristic polynomial $\psi_A(x) = f(x)$.

9.36 (i) Give an example of two nonisomorphic finite abelian groups having the same order and the same exponent.

(ii) Give an example of two nonsimilar matrices having the same characteristic polynomial and the same minimum polynomial.

9.3 JORDAN CANONICAL FORMS

If k is a finite field, then $\mathrm{GL}(n, k)$ is a finite group, and so every element in it has finite order. Consider the group-theoretic question: What is the order of A in $\mathrm{GL}(3, \mathbb{F}_7)$, where

$$A = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 4 \\ 0 & 1 & 3 \end{bmatrix}?$$

Of course, we can compute the powers $A^2, A^3, \ldots$; Lagrange's theorem guarantees there is some $n \geq 1$ with $A^n = E$, but this procedure for finding the order of A is rather tedious. We recognize A as the companion matrix of

$$g(x) = x^3 - 3x^2 - 4x - 1 = x^3 - 3x^2 + 3x - 1 = (x-1)^3$$

(remember that $g(x) \in \mathbb{F}_7[x]$). Now A and PAP^{-1} are conjugates in the group $\mathrm{GL}(n, k)$ and, hence, they have the same order. But the powers of a companion matrix are complicated (e.g., the square of a companion matrix is not a companion matrix). We now give a second canonical form whose powers are easily calculated, and we shall use it to compute the order of A later in this section.

Definition. A 1×1 ***Jordan block*** is a matrix $J(\alpha, 1) = [\alpha]$. If $s \geq 2$, then an $s \times s$ ***Jordan block*** is a matrix $J(\alpha, s)$ of the form

$$J(\alpha, s) = \begin{bmatrix} \alpha & 0 & 0 & \cdots & 0 & 0 \\ 1 & \alpha & 0 & \cdots & 0 & 0 \\ 0 & 1 & \alpha & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \alpha & 0 \\ 0 & 0 & 0 & \cdots & 1 & \alpha \end{bmatrix}.$$

.

Here is a more compact description of a Jordan block. Let L denote the $s \times s$ matrix having all entries 0 except for 1's on the subdiagonal just below the main diagonal. In this notation, a Jordan block $J(\alpha, s)$ has the form

$$J(\alpha, s) = \alpha I + L.$$

Let us regard L as a linear transformation on k^s. If $e_1, \ldots, e_s$ is the standard basis, then $Le_i = e_{i+1}$ if $i < s$ while $Le_s = 0$. It follows easily that the matrix L^2 is all 0's except for 1's on the second subdiagonal below the main diagonal; L^3 is all 0's except for 1's on the third subdiagonal; L^{s-1} has 1 in the $s, 1$ position, with 0's everywhere else, and $L^s = 0$.

Lemma 9.44. *If $J = J(\alpha, s) = \alpha I + L$ is an $s \times s$ Jordan block, then for all $m \geq 1$,*

$$J^m = \alpha^m I + \sum_{i=1}^{s-1} \binom{m}{i} \alpha^{m-i} L^i.$$

Proof. Since L and αI commute (actually, αI commutes with every matrix), the collection of all linear combinations of the powers of αI and the powers of L is a (commutative) ring, and so the binomial theorem applies. Finally, note that all terms involving L^i for $i \geq s$ are 0 because $L^s = 0$. •

Example 9.45.
Different powers of L are "disjoint"; that is, if $m \neq n$ and the ij entry of L^n is nonzero, then the ij entry of L^m is zero:

$$\begin{bmatrix} \alpha & 0 \\ 1 & \alpha \end{bmatrix}^m = \begin{bmatrix} \alpha^m & 0 \\ m\alpha^{m-1} & \alpha^m \end{bmatrix}$$

and

$$\begin{bmatrix} \alpha & 0 & 0 \\ 1 & \alpha & 0 \\ 0 & 1 & \alpha \end{bmatrix}^m = \begin{bmatrix} \alpha^m & 0 & 0 \\ m\alpha^{m-1} & \alpha^m & 0 \\ \binom{m}{2}\alpha^{m-2} & m\alpha^{m-1} & \alpha^m \end{bmatrix}. \quad ◀$$

Lemma 9.46. *If $g(x) = (x - \alpha)^s$, then the companion matrix $C(g)$ is similar to the $s \times s$ Jordan block $J(\alpha, s)$.*

Proof. If $T : k^s \to k^s$ is defined by $z \mapsto C(g)z$, then the proof of Lemma 9.33 gives a basis of k^s of the form $v, Tv, T^2v, \ldots, T^{s-1}v$. We claim that the list $Y = y_0, \ldots, y_{s-1}$ is also a basis of k^s, where

$$y_0 = v, y_1 = (T - \alpha I)v, \ldots, y_{s-1} = (T - \alpha I)^{s-1}v.$$

It is easy to see that the list Y spans V, because $T^i v \in \langle y_0, \ldots, y_i \rangle$ for all $0 \leq i \leq s - 1$. Since there are s elements in Y, Proposition 3.87 shows that Y is a basis.

We now compute $J = {}_Y[T]_Y$. If $j + 1 \leq s$, then

$$\begin{aligned} Ty_j &= T(T - \alpha I)^j v \\ &= (T - \alpha I)^j Tv \\ &= (T - \alpha I)^j [\alpha I + (T - \alpha I)]v \\ &= \alpha(T - \alpha I)^j v + (T - \alpha I)^{j+1} v. \end{aligned}$$

Hence, if $j + 1 < s$, then

$$Ty_j = \alpha y_j + y_{j+1}.$$

If $j + 1 = s$, then

$$(T - \alpha I)^{j+1} v = (T - \alpha I)^s v = 0,$$

by the Cayley–Hamilton theorem [for $\psi_{C(g)}(x) = (x - \alpha)^s$]; hence, $Ty_{s-1} = \alpha y_{s-1}$. The matrix J is thus a Jordan block $J(\alpha, s)$. By Corollary 3.101, $C(g)$ and $J(\alpha, s)$ are similar. •

It follows that Jordan blocks, as companion matrices, correspond to polynomials; in particular, $J(\alpha, s)$ corresponds to $(x-\alpha)^s$.

Theorem 9.47. *Let A be an $n \times n$ matrix with entries in a field k. If k contains all the eigenvalues of A (in particular, if k is algebraically closed), then A is similar to a direct sum of Jordan blocks.*

Proof. Instead of using the invariant factors $g_1 \mid g_2 \mid \cdots \mid g_t$, we are now going to use the elementary divisors $f_i(x)$ occurring in the basis theorem itself; that is, each $f_i(x)$ is a power of an irreducible polynomial in $k[x]$. By Theorem 9.34(i), a decomposition of $(k^n)^A$ into a direct sum of cyclic $k[x]$-modules W_i yields a direct sum of companion matrices

$$U = C(f_1) \oplus \cdots \oplus C(f_r),$$

where (f_i) is the order ideal of W_i, and U is similar to A. Since $\psi_A(x) = \prod_i f_i(x)$, however, our hypothesis says that each $f_i(x)$ splits over k; that is, $f_i(x) = (x-\alpha_i)^{s_i}$ for some $s_i \geq 1$, where α_i is an eigenvalue of A. By the lemma, $C(f_i)$ is similar to a Jordan block and, by Exercise 9.34 on page 674, A is similar to a direct sum of Jordan blocks. •

Definition. A ***Jordan canonical form*** is a direct sum of Jordan blocks.

If a matrix A is similar to the Jordan canonical form

$$J(\alpha_1, s_1) \oplus \cdots \oplus J(\alpha_r, s_r),$$

then we say that A has ***elementary divisors*** $(x-\alpha_1)^{s_1}, \ldots, (x-\alpha_r)^{s_r}$.

Theorem 9.47 says that every square matrix A having entries in a field containing all the eigenvalues of A is similar to a Jordan canonical form. Can a matrix be similar to several Jordan canonical forms? The answer is yes, but not really.

Example 9.48.
Let I_r be the $r \times r$ identity matrix, and let I_s be the $s \times s$ identity matrix. Then interchanging blocks in a direct sum yields a similar matrix:

$$\begin{bmatrix} B & 0 \\ 0 & A \end{bmatrix} = \begin{bmatrix} 0 & I_r \\ I_s & 0 \end{bmatrix} \begin{bmatrix} A & 0 \\ 0 & B \end{bmatrix} \begin{bmatrix} 0 & I_s \\ I_r & 0 \end{bmatrix}$$

Since every permutation is a product of transpositions, it follows that permuting the blocks of a matrix of the form $A_1 \oplus A_2 \oplus \cdots \oplus A_t$ yields a matrix similar to the orginal one. ◀

Theorem 9.49. *If A and B are $n \times n$ matrices over a field k containing all their eigenvalues, then A and B are similar if and only if they have the same elementary divisors. Moreover, if a matrix A is similar to two Jordan canonical forms, say, H and H', then H and H' have the same Jordan blocks (i.e., H' arises from H by permuting its Jordan blocks).*

Proof. By Corollary 3.101, A and B are similar if and only if $(k^n)^A \cong (k^n)^B$. By Theorem 9.22, $(k^n)^A \cong (k^n)^B$ if and only if their elementary divisors are the same.

In contrast to the invariant factors, which are given in a specific order (each dividing the next), A determines only a *set* of elementary divisors, hence only a set of Jordan blocks. By Example 9.48, the different Jordan canonical forms obtained from a given Jordan canonical form by permuting its Jordan blocks are all similar. •

Here are some applications of the Jordan canonical form.

Proposition 9.50. *If A is an $n \times n$ matrix with entries in a field k, then A is similar to its transpose A^t.*

Proof. First, Corollary 9.36(ii) allows us to assume that k contains all the eigenvalues of A. Now if $B = PAP^{-1}$, then $B^t = (P^t)^{-1}A^tP^t$; that is, if B is similar to A, then B^t is similar to A^t. Thus, it suffices to prove that H is similar to H^t for a Jordan canonical form H, and, by Exercise 9.34 on page 674, it is enough to show that a Jordan block $J = J(\alpha, s)$ is similar to J^t.

We illustrate the idea for $J(\alpha, 3)$. Let Q be the matrix having 1's on the "wrong" diagonal and 0's everywhere else; notice that $Q = Q^{-1}$.

$$\begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} \alpha & 0 & 0 \\ 1 & \alpha & 0 \\ 0 & 1 & \alpha \end{bmatrix} \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix} = \begin{bmatrix} \alpha & 1 & 0 \\ 0 & \alpha & 1 \\ 0 & 0 & \alpha \end{bmatrix}$$

We let the reader prove, in general, that $Q = Q^{-1}$ and $QJ(\alpha, s)Q^{-1} = J(\alpha, s)^t$. Perhaps the most efficient proof is to let $v_1, \ldots, v_s$ be a basis of a vector space W, to define $Q\colon W \to W$ by $Q\colon v_i \mapsto v_{s-i+1}$, and to define $J\colon W \to W$ by $J\colon v_i \mapsto \alpha v_i + v_{i+1}$ for $i < s$ and $J\colon v_s \mapsto \alpha v_s$. •

Example 9.51.
At the beginning of this section, we asked for the order of the matrix A in $\mathrm{GL}(3, \mathbb{F}_7)$, where

$$A = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 4 \\ 0 & 1 & 3 \end{bmatrix};$$

we saw that A is the companion matrix of $(x-1)^3$. Since $\psi_A(x)$ is a power of $x - 1$, the eigenvalues of A are all equal to 1 and hence lie in $\mathbb{F}_7$; by Lemma 9.46, A is similar to the Jordan block

$$J = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \end{bmatrix}.$$

By Example 9.45,

$$J^m = \begin{bmatrix} 1 & 0 & 0 \\ m & 1 & 0 \\ \binom{m}{2} & m & 1 \end{bmatrix},$$

and it follows that $J^7 = I$ because, in $\mathbb{F}_7$, we have $7 = 0$ and $\binom{7}{2} = 21 = 0$. Hence, A has order 7 in $\mathrm{GL}(3, \mathbb{F}_7)$. ◄

Exponentiating a matrix is used to find solutions of systems of linear differential equations; it is also very useful in setting up the relation between a Lie group and its corresponding Lie algebra. An $n \times n$ complex matrix A consists of n^2 entries, and so A may be regarded as a point in $\mathbb{C}^{n^2}$. This allows us to define convergence of a sequence of $n \times n$ complex matrices: $A_1, A_2, \ldots, A_k, \ldots$ ***converges*** to a matrix M if, for each i, j, the sequence of i, j entries converges. As in calculus, convergence of a series means convergence of the sequence of its partial sums.

Definition. If A is an $n \times n$ complex matrix, then

$$e^A = \sum_{k=0}^{\infty} \frac{1}{k!} A^k = I + A + \tfrac{1}{2}A^2 + \tfrac{1}{6}A^3 + \cdots .$$

It can be proved that this series converges for every matrix A, and that the function $A \mapsto e^A$ is continuous; that is, if $\lim_{k\to\infty} A_k = M$, then

$$\lim_{k\to\infty} e^{A_k} = e^M.$$

Here are some properties of this exponentiation of matrices; we shall see that the Jordan canonical form allows us to compute e^A.

Proposition 9.52. *Let A be an $n \times n$ complex matrix.*

(i) *If P is nonsingular, then $Pe^AP^{-1} = e^{PAP^{-1}}$.*

(ii) *If $AB = BA$, then $e^Ae^B = e^{A+B}$.*

(iii) *For every matrix A, the matrix e^A is nonsingular*; *indeed,*

$$(e^A)^{-1} = e^{-A}.$$

(iv) *If L is an $n \times n$ matrix having 1's just below the main diagonal and 0's elsewhere, then e^L is a lower triangular matrix with 1's on the diagonal.*

(v) *If D is a diagonal matrix, say, $D = \mathrm{diag}(\alpha_1, \alpha_2, \ldots, \alpha_n)$, then*

$$e^D = \mathrm{diag}(e^{\alpha_1}, e^{\alpha_2}, \ldots, e^{\alpha_n}).$$

(vi) *If $\alpha_1, \ldots, \alpha_n$ are the eigenvalues of A (with multiplicities), then $e^{\alpha_1}, \ldots, e^{\alpha_n}$ are the eigenvalues of e^A.*

(vii) *We can compute e^A.*

(viii) *If $\mathrm{tr}(A) = 0$, then $\det(e^A) = 1$.*

Proof. (i) We use the continuity of matrix exponentiation.

$$\begin{aligned} Pe^{A}P^{-1} &= P\Big(\lim_{n\to\infty}\sum_{k=0}^{n}\frac{1}{k!}A^{k}\Big)P^{-1} \\ &= \lim_{n\to\infty}\sum_{k=0}^{n}\frac{1}{k!}\big(PA^{k}P^{-1}\big) \\ &= \lim_{n\to\infty}\sum_{k=0}^{n}\frac{1}{k!}\big(PAP^{-1}\big)^{k} \\ &= e^{PAP^{-1}} \end{aligned}$$

(ii) The coefficient of the kth term of the power series for e^{A+B} is

$$\frac{1}{k!}(A+B)^{k},$$

while the coefficient of the kth term of $e^{A}e^{B}$ is

$$\sum_{i+j=k}\frac{1}{i!}A^{i}\frac{1}{j!}B^{j} = \sum_{i=0}^{k}\frac{1}{i!(k-i)!}A^{i}B^{k-i} = \frac{1}{k!}\sum_{i=0}^{k}\binom{k}{i}A^{i}B^{k-i}.$$

Since A and B commute, the binomial theorem shows that both kth coefficients are equal. See Exercise 9.44 on page 682 for an example where this is false if A and B do not commute.

(iii) This follows immediately from part (ii), for $-A$ and A commute and $e^{0} = I$.

(iv) The equation

$$e^{L} = I + L + \tfrac{1}{2}L^{2} + \cdots + \tfrac{1}{(s-1)!}L^{s-1}$$

holds because $L^{s} = 0$, and the result follows by Lemma 9.44. For example, when $s = 5$,

$$e^{L} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ \frac{1}{2} & 1 & 1 & 0 & 0 \\ \frac{1}{6} & \frac{1}{2} & 1 & 1 & 0 \\ \frac{1}{24} & \frac{1}{6} & \frac{1}{2} & 1 & 1 \end{bmatrix}.$$

(v) This is clear from the definition:

$$e^{D} = I + D + \tfrac{1}{2}D^{2} + \tfrac{1}{6}D^{3} + \cdots,$$

for

$$D^k = \operatorname{diag}(\alpha_1^k, \alpha_2^k, \ldots, \alpha_n^k).$$

(vi) Since $\mathbb{C}$ is algebraically closed, A is similar to its Jordan canonical form J: There is a nonsingular matrix P with $PAP^{-1} = J$. Now A and J have the same characteristic polynomials, and hence the same eigenvalues with multiplicities. But J is a lower triangular matrix with the eigenvalues $\alpha_1, \ldots, \alpha_n$ of A on the diagonal, and so the definition of matrix exponentiation gives e^J lower triangular with $e^{\alpha_1}, \ldots, e^{\alpha_n}$ on the diagonal. Since $e^A = e^{P^{-1}JP} = P^{-1}e^J P$, it follows that the eigenvalues of e^A are as claimed.

(vii) By Exercise 9.38, there is a nonsingular matrix P with $PAP^{-1} = \Delta + L$, where Δ is a diagonal matrix, $L^n = 0$, and $\Delta L = L\Delta$. Hence,

$$Pe^A P^{-1} = e^{PAP^{-1}} = e^{\Delta + L} = e^{\Delta} e^L.$$

But e^{Δ} is computed in part (v) and e^L is computed in part (iv). Hence, $e^A = P^{-1}e^{\Delta}e^L P$ is computable.

(viii) By definition, the trace of a matrix is the sum of its eigenvalues, while the determinant of a matrix is the product of the eigenvalues. Since the eigenvalues of e^A are $e^{\alpha_1}, \ldots, e^{\alpha_n}$, we have

$$\det(e^A) = \prod_i e^{\alpha_i} = e^{\sum_i \alpha_i} = e^{\operatorname{tr}(A)}.$$

Hence, $\operatorname{tr}(A) = 0$ implies $\det(e^A) = 1$. •

EXERCISES

9.37 Find all $n \times n$ matrices A over a field k for which A and A^2 are similar.

9.38 (**Jordan Decomposition**)
Prove that every $n \times n$ matrix A over an algebraically closed field k can be written

$$A = D + N,$$

where D is ***diagonalizable*** (i.e., D is similar to a diagonal matrix), N is ***nilpotent*** (i.e., $N^m = 0$ for some $m \geq 1$), and $DN = ND$.

9.39 Give an example of an $n \times n$ matrix that is not diagonalizable.
Hint. It is known that every symmetric real matrix is diagonalizable. Alternatively, a rotation (not the identity) about the origin on $\mathbb{R}^2$ sends no line through the origin into itself.

9.40 (i) Prove that all the eigenvalues of a nilpotent matrix are 0.
(ii) Use the Jordan form to prove the converse: If all the eigenvalues of a matrix A are 0, then A is nilpotent. (This result also follows from the Cayley–Hamilton theorem.)

9.41 How many similarity classes of 6×6 nilpotent real matrices are there?

9.42 If A is a nonsingular matrix and A is similar to B, prove that A^{-1} is similar to B^{-1}.

9.43 (i) Prove that every nilpotent matrix N is similar to a strictly lower triangular matrix (i.e., all entries on and above the diagonal are 0).

(ii) If N is a nilpotent matrix, prove that $I + N$ is nonsingular.

9.44 Let

$$A = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \text{ and } B = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}.$$

Prove that $e^A e^B \neq e^B e^A$, and conclude that $e^A e^B \neq e^{A+B}$.

9.45 How many conjugacy classes are there in $\mathrm{GL}(3, \mathbb{F}_7)$?

9.46 We know that $\mathrm{PSL}(3, \mathbb{F}_4)$ is a simple group of order $20\,160 = \frac{1}{2}8!$. Now A_8 contains an element of order 15, namely, $(1\ 2\ 3\ 4\ 5)(6\ 7\ 8)$. Prove that $\mathrm{PSL}(3, \mathbb{F}_4)$ has no element of order 15, and conclude that $\mathrm{PSL}(3, \mathbb{F}_4) \ncong A_8$.

Hint. Use Corollary 9.36 to replace $\mathbb{F}_4$ by a larger field containing any needed eigenvalues of a matrix. Compute the order [in the group $\mathrm{PSL}(3, \mathbb{F}_4)$] of the possible Jordan canonical forms

$$A = \begin{bmatrix} a & 0 & 0 \\ 1 & a & 0 \\ 0 & 1 & a \end{bmatrix}, \ B = \begin{bmatrix} a & 0 & 0 \\ 0 & b & 0 \\ 0 & 1 & b \end{bmatrix}, \text{ and } C = \begin{bmatrix} a & 0 & 0 \\ 0 & b & 0 \\ 0 & 0 & c \end{bmatrix}.$$

9.4 SMITH NORMAL FORMS

There is a defect in our account of canonical forms: How do we find the invariant factors of a given matrix A? The coming discussion will give an algorithm for computing its invariant factors. In particular, it will compute the minimum polynomial of A.

In Chapter 3, we showed that a linear transformation $T: V \to W$ between finite-dimensional vector spaces determines a matrix, once bases Y of V and Z of W are chosen, and Proposition 3.98 shows that matrix multiplication arises as the matrix determined by the composite of two linear transformations. We now generalize that calculation to R-maps between free R-modules, where R is any commutative ring.

Definition. Let R be a commutative ring, and let $T: R^t \to R^n$ be an R-map, where R^t and R^n are free R-modules of ranks t and n, respectively. If $Y = y_1, \ldots, y_t$ is a basis of R^t and $Z = z_1, \ldots, z_n$ is a basis of R^n, then

$$_Z[T]_Y = [a_{ij}]$$

is the $n \times t$ matrix over R whose ith column, for all i, consists of the coordinates of $T(y_i)$; that is,

$$T(y_i) = \sum_{j=1}^{n} a_{ji} z_j.$$

We now compare matrices for an R-homomorphism T arising from different choices of bases in R^t and in R^n. The next proposition generalizes Corollary 3.101 from vector spaces to modules over a commutative ring.

Proposition 9.53. *Let R be a commutative ring, let R^t and R^n be free R-modules of ranks t and n, respectively, and let $T\colon R^t \to R^n$ be an R-homomorphism. Let Y and Y' be bases of R^t, and let Z and Z' be bases of R^n. If $\Gamma = {}_Z[T]_Y$ and $\Gamma' = {}_{Z'}[T]_{Y'}$, then there exist invertible matrices P and Q, where P is $t \times t$ and Q is $n \times n$, with*

$$\Gamma' = Q\Gamma P^{-1}.$$

Conversely, if Γ and Γ' are $n \times t$ matrices over R with $\Gamma' = Q\Gamma P^{-1}$ for some invertible matrices P and Q, then there is an R-homomorphism $T\colon R^t \to R^n$, bases Y and Y' of R^t, and bases Z and Z' of R^n, respectively, such that $\Gamma = {}_Z[T]_Y$ and $\Gamma' = {}_{Z'}[T]_{Y'}$.

Proof. This is the same calculation we did in Proposition 3.101 on page 176 when we applied the formula

$${}_Z[S]_Y\,{}_Y[T]_X = {}_Z[ST]_X,$$

where $T\colon V \to V'$ and $S\colon V' \to V''$ and X, Y, and Z are bases of V, V', and V'', respectively. Note that the original proof never uses the inverse of any matrix entry, so that the earlier hypothesis that the entries lie in a field is much too strong; they may lie in any commutative ring. •

Definition. Two $n \times t$ matrices Γ and Γ' with entries in a commutative ring R are ***R-equivalent*** if there are invertible matrices P and Q with entries in R with

$$\Gamma' = Q\Gamma P$$

(writing P is just as general as writing P^{-1}).

Of course, equivalence as just defined is an equivalence relation on the set of all (rectangular) $n \times t$ matrices over R.

Proposition 9.54. *If R is a commutative ring, then finite presentations of* (*finitely presented*) *R-modules M and M' give exact sequences*

$$R^t \xrightarrow{\lambda} R^n \xrightarrow{\pi} M \to 0 \quad \text{and} \quad R^{t'} \xrightarrow{\lambda'} R^{n'} \xrightarrow{\pi'} M' \to 0,$$

and choices of bases Y, Y' of R^t and Z, Z' of R^n give matrices $\Gamma = {}_Z[\lambda]_Y$ and $\Gamma' = {}_{Z'}[\lambda']_{Y'}$. If $t' = t$, $n' = n$, and Γ and Γ' are R-equivalent, then $M \cong M'$.

Proof. Since Γ and Γ' are R-equivalent, there are invertible matrices P and Q with $\Gamma' = Q\Gamma P^{-1}$; now P determines an R-isomorphism $\theta\colon R^n \to R^n$, and Q determines an R-isomorphism $\varphi\colon R^t \to R^t$. The equation $\Gamma' = Q\Gamma P^{-1}$ implies that the following diagram commutes:

$$\begin{array}{ccccccc}
R^t & \xrightarrow{\lambda} & R^n & \xrightarrow{\pi} & M & \longrightarrow & 0 \\
\downarrow{\scriptstyle\varphi} & & \downarrow{\scriptstyle\theta} & & \downarrow{\scriptstyle\nu} & & \\
R^t & \xrightarrow[\lambda']{} & R^n & \xrightarrow[\pi']{} & M' & \longrightarrow & 0
\end{array}$$

Define an R-map $\nu\colon M \to M'$ as follows: if $m \in M$, then surjectivity of π gives an element $u \in R^n$ with $\pi(u) = m$; set $\nu(m) = \pi'\theta(u)$. Proposition 8.93 is a proof by diagram chasing that ν is a well-defined isomorphism. •

Proposition 9.54 is virtually useless; for most commutative rings R, there is no way to determine whether matrices Γ and Γ' with entries in R are R-equivalent. However, when R is a euclidean ring, we will be able to use the criterion in the proposition to find a computable normal form of a matrix.

If $T\colon V \to V$ is a linear transformation on a vector space V over a field k, the next theorem gives a finite presentation of the $k[x]$-module V^T. The next definition creates a free $k[x]$-module from any vector space V over a field k; the construction is based on the formal definition, in Chapter 3, of $k[x]$ as sequences in k almost all of whose coordinates are zero.

Definition. If V is a k-module over a commutative ring k, define

$$V[x] = \sum_{i \geq 0} V_i,$$

where $V_i \cong V$ for all i. In more detail, we denote the elements of V_i by $x^i v$, where $v \in V$ (so that x^i merely marks the coordinate position; in particular, we let $x^0 v = v$, so that $V_0 = V$). Thus, each element $u \in V[x]$ has a unique expression of the form

$$u = \sum_{i \geq 0} x^i v_i,$$

where $v_i \in V$ and almost all $v_i = 0$. The k-module $V[x]$ is a $k[x]$-module if we define

$$x\Big(\sum_i x^i v_i\Big) = \sum_i x^{i+1} v_i.$$

For readers comfortable with tensor product, the module $V[x]$ just constructed is merely $k[x] \otimes_k V$. Indeed, the next lemma uses the fact that tensor product commutes with direct sums, for a subset B is a basis of V if and only if $V = \sum_{b \in B} kb$ is a direct sum.

Lemma 9.55. *If V is a free k-module over a commutative ring k, then $V[x]$ is a free $k[x]$-module. In fact, a basis E of V is also a basis of $V[x]$ as a free $k[x]$-module.*

Proof. Each element $u \in V[x]$ has an expression of the form $u = \sum_{i \geq 0} x^i v_i$. Since $x^i e_1, \ldots, x^i e_n$ is a basis for $V_i = x^i V$, each $v_i = \sum_j \alpha_{ji} e_j$ for $\alpha_{ji} \in k$. Collecting terms,

$$u = f_1(x)e_1 + \cdots + f_n(x)e_n,$$

where $f_j(x) = \alpha_{j0} + \alpha_{j1}x + \cdots + \alpha_{jt}x^t$ for some t.

To prove uniqueness of this expression, suppose that

$$g_1(x)e_1 + \cdots + g_n(x)e_n = 0,$$

where $g_j(x) = \beta_{j0} + \beta_{j1}x + \cdots + \beta_{jt}x^t$ for some t. For each i, this gives the equation $\sum_j \beta_{ji}x^i e_j = 0$ in V_i. Since $x^i e_1, \ldots, x^i e_n$ is a basis of V_i, it is linearly independent, and so all $\beta_{ji} = 0$. •

We can now give a finite presentation of V^T. Viewing $V[x]$ as sequences (rather than as $k[x] \otimes_k V$) is convenient in this proof.

Theorem 9.56 (Characteristic Sequence).

(i) *If V is a finitely generated k-module over a commutative ring k and $T\colon V \to V$ is a k-homomorphism, then there is an exact sequence of $k[x]$-modules*

$$0 \to V[x] \xrightarrow{\lambda} V[x] \xrightarrow{\pi} V^T \to 0,$$

where, for all $i \geq 0$ and all $v \in V$, $\lambda(x^i v) = x^{i+1}v - x^i Tv$ and $\pi(x^i v) = T^i v$.

(ii) *If A is an $n \times n$ matrix over k and E is the standard basis $E = e_1, \ldots, e_n$ of k^n, then the matrix ${}_E[\lambda]_E$ arising from the presentation of $(k^n)^A$ in part* (i) *is $xI - A$.*

Proof. (i) It is easily checked that both λ and π are well-defined k-maps; they are also $k[x]$-maps; for example,

$$\lambda(x(x^i v)) = x\lambda(x^i v),$$

because both equal $x^{i+2}v - x^{i+1}Tv$.

(1) π is surjective. If $v \in V^T$, then $\pi(v) = T^0 v = v$.

(2) $\operatorname{im}\lambda \subseteq \ker\pi$.

$$\pi\lambda(x^i v) = \pi(x^{i+1}v - x^i Tv) = T^{i+1}v - T^{i+1}v = 0.$$

(3) $\ker\pi \subseteq \operatorname{im}\lambda$. If $u = \sum_{i=0}^m x^i v_i \in \ker\pi$, then $\sum_{i=0}^m T^i v_i = 0$. Hence,

$$\begin{aligned} u &= \sum_{i=0}^m x^i v_i - \sum_{i=0}^m T^i v_i \\ &= \sum_{i=1}^m (x^i v_i - T^i v_i), \end{aligned}$$

because

$$x^0 v_0 - T^0 v_0 = v_0 - v_0 = 0.$$

For any $i \geq 1$, we are going to rewrite the ith summand $x^i v_i - T^i v_i$ of u as a telescoping

sum, each of whose terms lies in $\operatorname{im}\lambda$; this will suffice to prove that $\ker\pi \subseteq \operatorname{im}\lambda$.

$$\begin{aligned}\sum_{j=0}^{i-1}\lambda(x^{i-1-j}T^j v_i) &= \sum_{j=0}^{i-1}\Big(x^{i-j}T^j v_i - x^{i-1-j}T^{j+1}v_i\Big)\\ &= (x^i v_i - x^{i-1}Tv_i) + (x^{i-1}Tv_i - x^{i-2}T^2 v_i) + \\ &\qquad \cdots + (xT^{i-1}v_i - T^i v_i)\\ &= x^i v_i + \Big[\sum_{j=1}^{i-1}(-x^{i-j}T^j v_i + x^{i-j}T^j v_i)\Big] - T^i v_i\\ &= x^i v_i - T^i v_i.\end{aligned}$$

(4) λ is injective. As a k-module, $V[x]$ is a direct sum of submodules V_i, and, for all $m \geq 0$, $V_m \cong V$ via $f_m \colon x^m v \mapsto v$; it follows that if $x^m v \neq 0$, then $f_{m+1}^{-1} f_m(x^m v) = x^{m+1}v \neq 0$.

Suppose now that

$$u = \sum_{i=0}^{m} x^i v_i \in \ker\lambda,$$

where $x^m v_m \neq 0$; it follows that $x^{m+1}v_m \neq 0$. But

$$0 = \lambda(u) = \lambda\Big(\sum_{i=0}^{m} x^i v_i\Big) = \sum_{i=0}^{m}(x^{i+1}v_i - x^i T v_i).$$

Therefore,

$$x^{m+1}v_m = -\sum_{i=0}^{m-1}(x^{i+1}v_i) + \sum_{i=0}^{m} x^i T v_i.$$

Thus, $x^{m+1}v_m \in V_{m+1} \cap \sum_{i=0}^{m} V_i = \{0\}$, so that $x^{m+1}v_m = 0$. But we have seen that $x^m v \neq 0$ implies $x^{m+1}v_m \neq 0$, so that this contradiction gives $\ker\lambda = \{0\}$.

(ii) In the notation of part (i), let $V = k^n$ and let $T \colon k^n \to k^n$ be given by $v \mapsto Av$, where v is an $n \times 1$ column vector. If $e_1, \ldots, e_n$ is the standard basis of k^n, then $e_1, \ldots, e_n$ is a basis of the free $k[x]$-module $V[x]$, and so it suffices to find the matrix of λ relative to this basis. Now

$$\lambda(e_i) = xe_i - Te_i = xe_i - \sum_j a_{ji}e_j.$$

Since $[\delta_{ij}] = I$, where δ_{ij} is the Kronecker delta, we have

$$\begin{aligned}xe_i - \sum_j a_{ji}e_j &= \sum_j x\delta_{ji}e_j - \sum_j a_{ji}e_j\\ &= \sum_j (x\delta_{ji} - a_{ji})e_j.\end{aligned}$$

Therefore, the matrix of λ is $xI - A$. •

Corollary 9.57. *Two $n \times n$ matrices A and B over a field k are similar if and only if the matrices $\Gamma = xI - A$ and $\Gamma' = xI - B$ are $k[x]$-equivalent.*

Proof. If A is similar to B, there is a nonsingular matrix P with entries in k such that $B = PAP^{-1}$. But

$$P(xI - A)P^{-1} = xI - PAP^{-1} = xI - B,$$

because the scalar matrix xI commutes with P (it commutes with every matrix). Thus, $xI - A$ and $xI - B$ are $k[x]$-equivalent.

Conversely, suppose that the matrices $xI - A$ and $xI - B$ are $k[x]$-equivalent. By Theorem 9.56(ii), $(k^n)^A$ and $(k^n)^B$ are finitely presented $k[x]$-modules with presentations that give the matrices $xI - A$ and $xI - B$, respectively. Now Proposition 9.54 shows that $(k^n)^A \cong (k^n)^B$, and so A and B are similar, by Corollary 7.4. •

Corollary 9.57 reduces the question of similarity of matrices over a field k to a problem of equivalence of matrices over $k[x]$. Fortunately, ***Gaussian elimination***, a method for solving systems of linear equations whose coefficients lie in a field, can be adapted here. We now generalize the ingredients of Gaussian elimination from matrices over fields to matrices over arbitrary commutative rings.

In what follows, we denote the ith row of a matrix A by $\text{ROW}(i)$ and the jth column by $\text{COL}(j)$.

Definition. There are three ***elementary row operations*** on a matrix A with entries in a commutative ring R:

Type I: Multiply $\text{ROW}(j)$ by a unit $u \in R$.

Type II: Replace $\text{ROW}(i)$ by $\text{ROW}(i) + c_j\text{ROW}(j)$, where $j \neq i$ and $c_j \in R$.

Type III: Interchange $\text{ROW}(i)$ and $\text{ROW}(j)$.

There are analogous ***elementary column operations***.

Notice that an operation of type III (i.e., an interchange) can be accomplished by operations of the other two types. We indicate this schematically.

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \to \begin{bmatrix} a-c & b-d \\ c & d \end{bmatrix} \to \begin{bmatrix} a-c & b-d \\ a & b \end{bmatrix} \to \begin{bmatrix} -c & -d \\ a & b \end{bmatrix} \to \begin{bmatrix} c & d \\ a & b \end{bmatrix}$$

Definition. An ***elementary matrix*** is the matrix obtained from the identity matrix I by applying an elementary row[10] operation to it.

Thus, there are three types of elementary matrix. It is shown in elementary linear algebra courses (and it is easy to prove) that performing an elementary operation is the same as multiplying by an elementary matrix. In more detail, if L is an elementary matrix of type I, II, or III, then applying an elementary row operation of this type to a matrix A gives

[10] Applying elementary column operations to I gives the same collection of elementary matrices.

the matrix LA, whereas applying the corresponding elementary column operation to A gives the matrix AL. It is also easy to see that every elementary matrix is invertible, and its inverse is elementary of the same type. It follows that every product of elementary matrices is invertible.

Definition. If R is a commutative ring, then a matrix Γ' is ***Gaussian equivalent*** to a matrix Γ if there is a sequence of elementary row and column operations

$$\Gamma = \Gamma_0 \to \Gamma_1 \to \cdots \to \Gamma_r = \Gamma'.$$

Gaussian equivalence is an equivalence relation on the family of all $n \times t$ matrices over R.

It follows that if Γ' is Gaussian equivalent to Γ, then there are matrices P and Q, each a product of elementary matrices, with $\Gamma' = P\Gamma Q$. Recall that two matrices Γ' and Γ are *R-equivalent* if there are invertible matrices P and Q with $\Gamma' = P\Gamma Q$. It follows that if Γ' is Gaussian equivalent to Γ, then Γ' and Γ are R-equivalent. We shall see that the converse is true when R is euclidean.

Theorem 9.58 (Smith[11] Normal Form). *Every nonzero $n \times t$ matrix Γ with entries in a euclidean ring R is Gaussian equivalent to a matrix of the form*

$$\begin{bmatrix} \Sigma & 0 \\ 0 & 0 \end{bmatrix},$$

where $\Sigma = \operatorname{diag}(\sigma_1, \ldots, \sigma_q)$ and $\sigma_1 \mid \sigma_2 \mid \cdots \mid \sigma_q$ are nonzero (the lower blocks of 0's or the right blocks of 0's may not be present).

Proof. The proof is by induction on the number $n \geq 1$ of rows of Γ. If $\sigma \in R$, let $\partial(\sigma)$ denote its degree in the euclidean ring R. Among all the entries of matrices Gaussian equivalent to Γ, let σ_1 have the smallest degree, and let Δ be a matrix Gaussian equivalent to Γ that has σ_1 as an entry, say, in position k, ℓ.

We claim that $\sigma_1 \mid \eta_{kj}$ for all η_{kj} in $\text{ROW}(k)$ of Δ. Otherwise, there is $j \neq \ell$ and an equation $\eta_{kj} = \kappa\sigma_1 + \rho$, where $\partial(\rho) < \partial(\sigma_1)$. Adding $(-\kappa)\text{COL}(\ell)$ to $\text{COL}(j)$ gives a matrix Δ' having ρ as an entry. But Δ' is Gaussian equivalent to Γ and has an entry ρ whose degree is smaller than $\partial(\sigma_1)$, a contradiction. The same argument shows that σ_1 divides any entry in its column. We claim that σ_1 divides every entry of Δ'. Let a be an entry not in σ_1's row or column; schematically, we have $\begin{pmatrix} a & b \\ c & \sigma_1 \end{pmatrix}$, where $b = u\sigma_1$ and $c = v\sigma_1$. Replace $\text{ROW}(1)$ by $\text{ROW}(1) + (1-u)\text{ROW}(2) = (a + (1-u)c \;\; \sigma_1)$. As above, $\sigma_1 \mid a + (1-u)c$. Since $\sigma_1 \mid c$, we have $\sigma_1 \mid a$.

Let us return to Δ, a matrix Gaussian equivalent to Γ that contains σ_1 as an entry. By interchanges, there is a matrix Δ' that is Gaussian equivalent to Γ and that has σ_1 in the 1,1

[11] This theorem and the corresponding uniqueness result, soon to be proved, were found by H. J. S. Smith in 1861.

position. If η_{1j} is another entry in the first row, then $\eta_{1j} = \kappa_j \sigma_1$, and adding $(-\kappa_j)\text{COL}(1)$ to $\text{COL}(j)$ gives a new matrix whose $1, j$ entry is 0. Thus, the matrix Δ is Gaussian equivalent to a matrix having σ_1 in the 1,1 position and with 0's in the rest of the first row. This completes the base step $n = 1$ of the induction, for we have just shown that a nonzero $1 \times t$ matrix is Gaussian equivalent to $[\sigma_1\ 0 \ldots 0]$. Furthermore, since σ_1 divides all entries in the first column, Γ is Gaussian equivalent to a matrix having all 0's in the rest of the first column as well; thus, Γ is Gaussian equivalent to a matrix of the form

$$\begin{bmatrix} \sigma_1 & 0 \\ 0 & \Omega \end{bmatrix}.$$

By induction, the matrix Ω is Gaussian equivalent to a matrix

$$\begin{bmatrix} \Sigma' & 0 \\ 0 & 0 \end{bmatrix},$$

where $\Sigma' = \text{diag}(\sigma_2, \ldots, \sigma_q)$ and $\sigma_2 \mid \sigma_3 \mid \cdots \mid \sigma_q$. Hence, Γ is Gaussian equivalent to $\begin{bmatrix} \sigma_1 & 0 & 0 \\ 0 & \Sigma' & 0 \\ 0 & 0 & 0 \end{bmatrix}$. It remains to observe that $\sigma_1 \mid \sigma_2$; this follows from our initial remarks, for the ultimate matrix is Gaussian equivalent to Γ and contains σ_1 as an entry. •

Definition. The matrix $\begin{bmatrix} \Sigma & 0 \\ 0 & 0 \end{bmatrix}$ in the statement of the theorem is called a ***Smith normal form*** of Γ.

Thus, the theorem states that every nonzero (rectangular) matrix with entries in a euclidean ring R is Gaussian equivalent to a Smith normal form.

Corollary 9.59. *Let R be a euclidean ring.*

(i) *Every invertible $n \times n$ matrix Γ with entries in R is a product of elementary matrices.*

(ii) *Two matrices Γ and Γ' over R are R-equivalent if and only if they are Gaussian equivalent.*

Proof. (i) We now know that Γ is Gaussian equivalent to a Smith normal form $\left[\begin{smallmatrix} \Sigma & 0 \\ 0 & 0 \end{smallmatrix}\right]$, where Σ is diagonal. Since Γ is a (square) invertible matrix, there can be no blocks of 0's, and so Γ is Gaussian equivalent to Σ; that is, there are matrices P and Q that are products of elementary matrices such that

$$P \Gamma Q = \Sigma = \text{diag}(\sigma_1, \ldots, \sigma_n).$$

Hence, $\Gamma = P^{-1} \Sigma Q^{-1}$. Now the inverse of an elementary matrix is again elementary, so that P^{-1} and Q^{-1} are products of elementary matrices. Since Σ is invertible, $\det(\Sigma) = \sigma_1 \cdots \sigma_n$ is a unit in R. It follows that each σ_i is a unit, and so Σ is a product of n

elementary matrices [arising from the elementary operations of multiplying $\text{ROW}(i)$ by the unit σ_i].

(ii) It is always true that if Γ' and Γ are Gaussian equivalent, then they are R-equivalent, for if $\Gamma' = P\Gamma Q$, where P and Q are products of elementary matrices, then P and Q are invertible. Conversely, if Γ' is R-equivalent to Γ, then $\Gamma' = P\Gamma Q$, where P and Q are invertible, and part (i) shows that Γ' and Γ are Gaussian equivalent. •

There are examples showing that this proposition may be false for PIDs that are not euclidean.[12] Investigating this phenomenon was important in the beginnings of *Algebraic K-Theory* (see Milnor, *Introduction to Algebraic K-Theory*).

Theorem 9.60 (Simultaneous Bases). *Let R be a euclidean ring, let F be a finitely generated free R-module, and let S be a submodule of F. Then there exists a basis $z_1, \ldots, z_n$ of F and nonzero $\sigma_1, \ldots, \sigma_q$ in R, where $0 \leq q \leq n$, such that $\sigma_1 \mid \cdots \mid \sigma_q$ and $\sigma_1 z_1, \ldots, \sigma_q z_q$ is a basis of S.*

Proof. If $M = F/S$, then Corollary 9.4 shows that S is free of rank $\leq n$, and so

$$0 \to S \xrightarrow{\lambda} F \to M \to 0$$

is a presentation of M, where λ is the inclusion. Now any choice of bases of S and F associates a matrix Γ to λ (note that Γ may be rectangular). According to Proposition 9.53, there are new bases of S and F relative to which Γ is R-equivalent to a Smith normal form, and these new bases are as described in the theorem. •

Corollary 9.61. *Let Γ be an $n \times n$ matrix with entries in a euclidean ring R.*

(i) *If Γ is R-equivalent to a Smith normal form* $\operatorname{diag}(\sigma_1, \ldots, \sigma_q) \oplus 0$, *then those $\sigma_1, \ldots, \sigma_q$ that are not units are the invariant factors of Γ.*

(ii) *If* $\operatorname{diag}(\eta_1, \ldots, \eta_s) \oplus 0$ *is another Smith normal form of Γ, then $s = q$ and there are units u_i with $\eta_i = u_i\sigma_i$ for all i; that is, the diagonal entries are associates.*

Proof. (i) We may regard Γ as the matrix associated to an R-map $\lambda\colon R^n \to R^n$ relative to some choice of bases. Let $M = R^n/\operatorname{im}\lambda$. If $\operatorname{diag}(\sigma_1, \ldots, \sigma_q) \oplus 0$ is a Smith normal form of Γ, then there are bases $y_1, \ldots, y_n$ of R^n and $z_1, \ldots, z_n$ of R^n with $\lambda(y_1) = \sigma_1 z_1, \ldots, \lambda(y_q) = \sigma_q z_q$ and $\lambda(y_j) = 0$ for all y_j with $j > q$, if any. If σ_s is the first σ_i that is not a unit, then

$$M \cong R^{n-q} \oplus R/(\sigma_s) \oplus \cdots \oplus R/(\sigma_q),$$

a direct sum of cyclic modules for which $\sigma_s \mid \cdots \mid \sigma_q$. The fundamental theorem of finitely generated R-modules identifies $\sigma_s, \ldots, \sigma_q$ as the invariant factors of M.

(ii) Part (i) proves the essential uniqueness of the Smith normal form, for the invariant factors, being generators of order ideals, are only determined up to associates. •

[12] There is a version for general PIDs obtained by augmenting the collection of elementary matrices by *secondary* matrices; see Exercise 9.50 on page 694.

With a slight abuse of language, we may now speak of *the* Smith normal form of a matrix.

Corollary 9.62. *Two $n \times n$ matrices A and B over a field k are similar if and only if $xI - A$ and $xI - B$ have the same Smith normal form over $k[x]$.*

Proof. By Theorem 9.57, A and B are similar if and only if $xI - A$ is $k[x]$-equivalent to $xI - B$, and Corollary 9.61 shows that $xI - A$ and $xI - B$ are $k[x]$-equivalent if and only if they have the same Smith normal form. •

Corollary 9.63. *Let F be a finitely generated free abelian group, and let S be a subgroup of F having finite index; let $y_1, \ldots, y_n$ be a basis of F, let $z_1, \ldots, z_n$ be a basis of S, and let $A = [a_{ij}]$ be the $n \times n$ matrix with $z_i = \sum_j a_{ji} y_j$. Then*

$$[F : S] = |\det(A)|.$$

Proof. Changing bases of S and of F replaces A by a matrix B that is $\mathbb{Z}$-equivalent to it:

$$B = QAP,$$

where Q and P are invertible matrices with entries in $\mathbb{Z}$. Since the only units in $\mathbb{Z}$ are 1 and -1, we have $|\det(B)| = |\det(A)|$. In particular, if we choose B to be a Smith normal form, then $B = \operatorname{diag}(g_1, \ldots, g_n)$, and so $|\det(B)| = g_1 \cdots g_n$. But $g_1, \ldots, g_n$ are the invariant factors of F/S; by Corollary 5.30, their product is the order of F/S, which is the index $[F : S]$. •

We have not yet kept our promise to give an algorithm computing the invariant factors of a matrix with entries in a field.

Theorem 9.64. *Let $\Sigma = \operatorname{diag}(\sigma_1, \ldots, \sigma_q)$ be the diagonal block in the Smith normal form of a matrix Γ with entries in a euclidean ring R. If we define $d_i(\Gamma)$ by $d_0(\Gamma) = 1$ and, for $i > 0$,*

$$d_i(\Gamma) = \gcd(\text{ all } i \times i \text{ minors of } \Gamma),$$

then, for all $i \geq 1$,

$$\sigma_i = d_i(\Gamma)/d_{i-1}(\Gamma).$$

Proof. We are going to show that if Γ and Γ' are R-equivalent, then

$$d_i(\Gamma) \sim d_i(\Gamma')$$

for all i, where we write $a \sim b$ to denote a and b being associates in R. This will suffice to prove the theorem, for if Γ' is the Smith normal form of Γ whose diagonal block is $\operatorname{diag}(\sigma_1, \ldots, \sigma_q)$, then $d_i(\Gamma') = \sigma_1 \sigma_2 \cdots \sigma_i$. Hence,

$$\sigma_i(x) = d_i(\Gamma')/d_{i-1}(\Gamma') \sim d_i(\Gamma)/d_{i-1}(\Gamma).$$

By Proposition 9.59, it suffices to prove that

$$d_i(\Gamma) \sim d_i(L\Gamma) \quad \text{and} \quad d_i(\Gamma) \sim d_i(\Gamma L)$$

for every elementary matrix L. Indeed, it suffices to prove that $d_i(\Gamma L) \sim d_i(\Gamma)$, because $d_i(\Gamma L) = d_i([\Gamma L]^t) = d_i(L^t\Gamma^t)$ [the $i \times i$ submatrices of Γ^t are the transposes of the $i \times i$ submatrices of Γ; now use the facts that L^t is elementary and that, for every square matrix M, we have $\det(M^t) = \det(M)$].

As a final simplification, it suffices to consider only elementary operations of types I and II, for we have seen on page 687 that an operation of type III, interchanging two rows, can be accomplished using the other two types.

L is of type I.

If we multiply $\text{ROW}(\ell)$ of Γ by a unit u, then an $i \times i$ submatrix either remains unchanged or one of its rows is multiplied by u. In the first case, the minor, namely, its determinant, is unchanged; in the second case, the minor is changed by a unit. Therefore, every $i \times i$ minor of $L\Gamma$ is an associate of the corresponding $i \times i$ minor of Γ, and so $d_i(L\Gamma) \sim d_i(\Gamma)$.

L is of type II.

If L replaces $\text{ROW}(\ell)$ by $\text{ROW}(\ell) + r\,\text{ROW}(j)$, then only $\text{ROW}(\ell)$ of Γ is changed. Thus, an $i \times i$ submatrix of Γ either does not involve this row or it does. In the first case, the corresponding minor of $L\Gamma$ is unchanged; in the second case, it has the form $m + rm'$, where m and m' are $i \times i$ minors of Γ (for det is a multilinear function of the rows of a matrix). It follows that $d_i(\Gamma) \mid d_i(L\Gamma)$, for $d_i(\Gamma) \mid m$ and $d_i(\Gamma) \mid m'$. Since L^{-1} is also an elementary matrix of type II, this argument shows that $d_i(L^{-1}(L\Gamma)) \mid d_i(L\Gamma)$. Of course, $L^{-1}(L\Gamma) = \Gamma$, so that $d_i(\Gamma)$ and $d_i(L\Gamma)$ divide each other. As R is a domain, we have $d_i(L\Gamma) \sim d_i(\Gamma)$. •

Theorem 9.65. *There is an algorithm to compute the elementary divisors of any square matrix A with entries in a field k.*

Proof. By Corollary 9.62, it suffices to find a Smith normal form for $\Gamma = xI - A$ over the ring $k[x]$; by Corollary 9.61, the invariant factors of A are those diagonal entries which are not units.

There are two algorithms: compute $d_i(xI - A)$ for all i (of course, this is not a very efficient algorithm for large matrices); put $xI - A$ into Smith normal form using Gaussian elimination over $k[x]$. The reader should now have no difficulty in writing a program to compute the elementary divisors. •

Example 9.66.

Find the invariant factors, over $\mathbb{Q}$, of

$$A = \begin{bmatrix} 2 & 3 & 1 \\ 1 & 2 & 1 \\ 0 & 0 & -4 \end{bmatrix}.$$

We are going to use a combination of the two modes of attack: Gaussian elimination and gcd's of minors. Now

$$xI - A = \begin{bmatrix} x-2 & -3 & -1 \\ -1 & x-2 & -1 \\ 0 & 0 & x+4 \end{bmatrix}.$$

It is plain that $g_1 = 1$, for it is the gcd of all the entries of A, some of which are nonzero constants. Interchange ROW(1) and ROW(2), and then change sign in the top row to obtain

$$\begin{bmatrix} 1 & -x+2 & 1 \\ x-2 & -3 & -1 \\ 0 & 0 & x+4 \end{bmatrix}.$$

Add $-(x-2)$ROW(1) to ROW(2) to obtain

$$\begin{bmatrix} 1 & -x+2 & 1 \\ 0 & x^2-4x+1 & -x+1 \\ 0 & 0 & x+4 \end{bmatrix} \to \begin{bmatrix} 1 & 0 & 0 \\ 0 & x^2-4x+1 & -x+1 \\ 0 & 0 & x+4 \end{bmatrix}.$$

The gcd of the entries in the 2×2 submatrix

$$\begin{bmatrix} x^2-4x+1 & -x+1 \\ 0 & x+4 \end{bmatrix}$$

is 1, for $-x+1$ and $x+4$ are distinct irreducibles, and so $g_2 = 1$. We have shown that there is only one invariant factor of A, namely, $(x^2-4x+1)(x+4) = x^3 - 15x + 4$, and it must be the characteristic polynomial of A. It follows that the characteristic and minimal polynomials of A coincide, and Corollary 9.43 shows that the rational canonical form of A is

$$\begin{bmatrix} 0 & 0 & -4 \\ 1 & 0 & 15 \\ 0 & 1 & 0 \end{bmatrix}. \quad ◀$$

Example 9.67.

Find the abelian group G having generators a, b, c and relations

$$\begin{aligned} 7a + 5b + 2c &= 0 \\ 3a + 3b \qquad &= 0 \\ 13a + 11b + 2c &= 0. \end{aligned}$$

Using elementary operations over $\mathbb{Z}$, we find the Smith normal form of the matrix of relations:

$$\begin{bmatrix} 7 & 5 & 2 \\ 3 & 3 & 0 \\ 13 & 11 & 2 \end{bmatrix} \to \begin{bmatrix} 1 & 0 & 0 \\ 0 & 6 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

It follows that $G \cong (\mathbb{Z}/1\mathbb{Z}) \oplus (\mathbb{Z}/6\mathbb{Z}) \oplus (\mathbb{Z}/0\mathbb{Z})$. Simplifying, $G \cong \mathbb{I}_6 \oplus \mathbb{Z}$. ◀

EXERCISES

9.47 Find the invariant factors, over $\mathbb{Q}$, of the matrix

$$\begin{bmatrix} -4 & 6 & 3 \\ -3 & 5 & 4 \\ 4 & -5 & 3 \end{bmatrix}.$$

9.48 Find the invariant factors, over $\mathbb{Q}$, of the matrix

$$\begin{bmatrix} -6 & 2 & -5 & -19 \\ 2 & 0 & 1 & 5 \\ -2 & 1 & 0 & -5 \\ 3 & -1 & 2 & 9 \end{bmatrix}.$$

9.49 If k is a field, prove that there is an additive exact functor ${}_k\mathbf{Mod} \to {}_{k[x]}\mathbf{Mod}$ taking any vector space V to $V[x]$. [See Theorem 9.56(ii).]

9.50 Let R be a PID, and let $a, b \in R$.

(i) If d is the gcd of a and b, prove that there is a 2×2 matrix $Q = \begin{bmatrix} x & y \\ x' & y' \end{bmatrix}$ with $\det(Q) = 1$ so that

$$Q \begin{bmatrix} a & * \\ b & * \end{bmatrix} = \begin{bmatrix} d & * \\ d' & * \end{bmatrix},$$

where $d \mid d'$.

Hint. If $d = xa + yb$, define $x' = b/d$ and $y' = -a/d$.

(ii) Call an $n \times n$ matrix U ***secondary*** if it can be partitioned

$$U = \begin{bmatrix} Q & 0 \\ 0 & I \end{bmatrix},$$

where Q is a 2×2 matrix of determinant 1. Prove that every $n \times n$ matrix A with entries in a PID can be transformed into a Smith canonical form by a sequence of elementary and secondary matrices.

9.5 BILINEAR FORMS

In this section, k will be a field and V will be a vector space over k, usually finite-dimensional. Even though we have not yet proved the basic theorems about determinants (they will be proved in Section 9.9), we continue to use their familiar properties.

Definition. A ***bilinear form*** (or ***inner product***) on V is a bilinear function

$$f : V \times V \to k.$$

The ordered pair (V, f) is called an ***inner product space***.

Of course, (k^n, f) is an inner product space if f is the familiar ***dot product***

$$f(u, v) = \sum_i u_i v_i,$$

where $u = (u_1, \ldots, u_n)^t$ and $v = (v_1, \cdots, v_n)^t$ (the superscript t denotes transpose; remember that the elements of k^n are $n \times 1$ column vectors). In terms of matrix multiplication, we have

$$f(u, v) = u^t v.$$

There are two types of bilinear forms of special interest.

Definition. A bilinear form $f : V \times V \to k$ is ***symmetric*** if

$$f(u, v) = f(v, u)$$

for all $u, v \in V$; we call an inner product space (V, f) a ***symmetric space*** when f is symmetric.

A bilinear form f is ***alternating*** if $f(v, v) = 0$ for all $v \in V$; we call an inner product space (V, f) an ***alternating space*** when f is alternating.

Example 9.68.
(i) If $V = k^2$ and its elements are viewed as column vectors, then $\det : V \times V \to k$, given by

$$\left(\begin{bmatrix} a \\ b \end{bmatrix}, \begin{bmatrix} c \\ d \end{bmatrix} \right) \mapsto \det \begin{bmatrix} a & c \\ b & d \end{bmatrix} = ad - bc,$$

is an example of an alternating bilinear form.

(ii) In Chapter 8, we defined a (Hermitian) form on the complex vector space $\mathrm{cf}(G)$ of all class functions on a finite group G. More generally, define a function $f : \mathbb{C}^n \times \mathbb{C}^n \to \mathbb{C}$ by

$$f(u, v) = \sum_j u_j \overline{v}_j,$$

where $u = (u_1, \ldots, u_n)^t$, $v = (v_1, \ldots, v_n)^t$, and $\overline{c}$ denotes the complex conjugate of a complex number c. Such a function is not $\mathbb{C}$-bilinear because $f(u, cv) = \overline{c} f(u, v)$ instead of $cf(u, v)$. Hermitian forms are examples of *sesquilinear forms*; such forms can be constructed over any field k equipped with an automorphism of order 2 (to play the role of complex conjugation). ◂

Every bilinear form can be expressed in terms of symmetric and alternating bilinear forms.

Proposition 9.69. *Let k be a field of characteristic $\neq 2$, and let f be a bilinear form defined on a vector space V over k. Then there are unique bilinear forms f_s and f_a, where f_s is symmetric and f_a is alternating, such that $f = f_s + f_a$.*

Proof. By hypothesis, $\frac{1}{2} \in k$, and so we may define

$$f_s(u, v) = \tfrac{1}{2}\big(f(u, v) + f(v, u)\big)$$

and

$$f_a(u, v) = \tfrac{1}{2}\big(f(u, v) - f(v, u)\big).$$

It is clear that $f = f_s + f_a$, that f_s is symmetric, and that f_a is alternating. Let us prove uniqueness. If $f = f'_s + f'_a$, where f'_s is symmetric and f'_a is alternating, then $f_s + f_a = f'_s + f'_a$, so that $f_s - f'_s = f'_a - f_a$. If we define g to be the common value, $f_s - f'_s = g = f'_a - f_a$, then g is both symmetric and alternating. By Exercise 9.54 on page 713, we have $g = 0$, and so $f_s = f'_s$ and $f_a = f'_a$. •

Remark. If (V, g) is an inner product space, then g is called ***skew*** if

$$g(v, u) = -g(u, v)$$

for all $u, v \in V$. We now show that if k does not have characteristic 2, then g is alternating if and only if g is skew.

If g is any bilinear form, we have

$$g(u + v, u + v) = g(u, u) + g(u, v) + g(v, u) + g(v, v).$$

Therefore, if g is alternating, then $0 = g(u, v) + g(v, u)$, so that g is skew. (We have not yet used the hypothesis that the characteristic of k is not 2.)

Conversely, if g is skew, then set $u = v$ in the equation $g(u, v) = -g(v, u)$ to get $g(u, u) = -g(u, u)$; that is, $2g(u, u) = 0$. Thus, $g(u, u) = 0$, because k does not have characteristic 2, and so g is alternating. ◂

Definition. Let f be a bilinear form on a vector space V over a field k, and let $E = e_1, \ldots, e_n$ be a basis of V. Then an ***inner product matrix*** of f ***relative to*** E is

$$A = [f(e_i, e_j)].$$

Suppose that (V, f) is an inner product space, $e_1, \ldots, e_n$ is a basis of V, and $A = [f(e_i, e_j)]$ is the inner product matrix of f relative to E. If $b = \sum b_i e_i$ and $c = \sum c_i e_i$ are vectors in V, then

$$f(b, c) = f\big(\sum b_i e_i, \sum c_i e_i\big) = \sum_{i,j} b_i f(e_i, e_j) c_j.$$

If B and C denote the column vectors $(b_1, \ldots, b_n)^t$ and $(c_1, \ldots, c_n)^t$, respectively, then this last equation can be written in matrix form:

$$f(b, c) = B^t AC.$$

Thus, an inner product matrix determines f completely.

Proposition 9.70. *Let V be an n-dimensional vector space over a field k.*

(i) *Every $n \times n$ matrix A over a field k is the inner product matrix of some bilinear form f defined on $V \times V$. If f is symmetric, then its inner product matrix A relative to any basis of V is a symmetric matrix (i.e., $A^t = A$). If f is alternating, then any inner product matrix relative to any basis of V is skew-symmetric (i.e., $A^t = -A$).*

(ii) *If $B^t AC = B^t A'C$ for all column vectors B and C, then $A = A'$.*

(iii) *Let A and A' be inner product matrices of bilinear forms f and f' on V relative to bases E and E', respectively. Then $f = f'$ if and only if A and A' are* ***congruent****; that is, there exists a nonsingular matrix P with*

$$A' = P^t AP.$$

Proof. (i) For any matrix A, the function $f : k^n \times k^n \to k$, defined by $f(b, c) = b^t Ac$, is easily seen to be a bilinear form, and A is its inner product matrix relative to the standard basis $e_1, \ldots, e_n$. The reader may easily transfer this construction to any vector space V once a basis of V is chosen.

If f is symmetric, then so is its inner product matrix $A = [a_{ij}]$, for $a_{ij} = f(e_i, e_j) = f(e_j, e_i) = a_{ji}$; similarly, if f is alternating, then $a_{ij} = f(e_i, e_j) = -f(e_j, e_i) = -a_{ji}$.

(ii) If $b = \sum_i b_i e_i$ and $c = \sum_i c_i e_i$, then we have seen that $f(b, c) = B^t AC$, where B and C are the column vectors of the coordinates of b and c with respect to E. In particular, if $b = e_i$ and $c = e_j$, then $f(e_i, e_j) = a_{ij}$ is the ij entry of A.

(iii) Let the coordinates of b and c with respect to the basis E' be B' and C', respectively, so that $f'(b, c) = (B')^t A'C'$, where $A' = [f(e_i', e_j')]$. If P is the transition matrix ${}_E[1]_{E'}$, then $B = PB'$ and $C = PC'$. Hence, $f(b, c) = B^t AC = (PB')^t A(PC') = (B')^t (P^t AP)C'$. By part (ii), we must have $P^t AP = A'$.

For the converse, the given matrix equation $A' = P^t AP$ yields equations:

$$\begin{aligned} [f'(e_i', e_j')] &= A' \\ &= P^t AP \\ &= \Big[\sum_{\ell q} p_{\ell i} f(e_\ell, e_q) p_{qj})\Big] \\ &= \Big[f\Big(\sum_\ell p_{\ell i} e_\ell, \sum_q p_{qj} e_q\Big)\Big] \\ &= [f(e_i', e_j')]. \end{aligned}$$

Hence, $f'(e_i', e_j') = f(e_i', e_j')$ for all i, j, from which it follows that $f'(b, c) = f(b, c)$ for all $b, c \in V$. Therefore, $f = f'$. •

Corollary 9.71. *If (V, f) is an inner product space and if A and A' are inner product matrices of f relative to different bases of V, then there exists a nonzero $a \in k$ with*

$$\det(A') = a^2 \det(A).$$

Consequently, A' is nonsingular if and only if A is nonsingular.

Proof. This follows from the facts: $\det(P^t) = \det(P)$; $\det(AB) = \det(A)\det(B)$; and P is nonsingular if and only if $\det(P) \neq 0$. •

The most important bilinear forms are the *nondegenerate* ones.

Definition. A bilinear form f is ***nondegenerate*** if it has a nonsingular inner product matrix.

For example, the dot product on k^n is nondegenerate, for its inner product matrix relative to the standard basis is the identity matrix I.

The *discriminant* of a bilinear form is essentially the determinant of its inner product matrix. However, since the inner product matrix depends on a choice of basis, we must complicate the definition a bit.

Definition. If k is a field, then its multiplicative group of nonzero elements is denoted by $k^\times$. Define $(k^\times)^2 = \{a^2 : a \in k^\times\}$. The ***discriminant*** of a bilinear form f is either 0 or

$$\det(A)(k^\times)^2 \in k^\times/(k^\times)^2,$$

where A is an inner product matrix of f.

It follows from Corollary 9.71 that the discriminant of f is well-defined. Quite often, however, we are less careful and say that $\det(A)$ is the discriminant of f, where A is some inner product matrix of f.

The next definition will be used in characterizing nondegeneracy.

Definition. If (V, f) is an inner product space and $W \subseteq V$ is a subspace of V, then the ***left orthogonal complement*** of W is

$$W^{\perp L} = \{b \in V : f(b, w) = 0 \text{ for all } w \in W\};$$

the ***right orthogonal complement*** of W is

$$W^{\perp R} = \{c \in V : f(w, c) = 0 \text{ for all } w \in W\}.$$

It is easy to see that both $W^{\perp L}$ and $W^{\perp R}$ are subspaces of V. Moreover, $W^{\perp L} = W^{\perp R}$ if f is either symmetric or alternating, in which case we write $W^\perp$.

Let (V, f) be an inner product space, and let A be the inner product matrix of f relative to a basis $e_1, \ldots, e_n$ of V. We claim that $b \in V^{\perp L}$ if and only if b is a solution of the homogeneous system $A^t x = 0$. If $b \in V^{\perp L}$. then $f(b, e_j) = 0$ for all j. Writing $b = \sum_i b_i e_i$, we see that $0 = f(b, e_j) = f\big(\sum_i b_i e_i, e_j\big) = \sum_j b_i f(e_i, e_j)$. In matrix terms, $b = (b_1, \ldots, b_n)^t$ and $B^t A = 0$; transposing, b is a solution of the homogeneous system $A^t x = 0$. The proof of the converse is left to the reader. A similar argument shows that $c \in V^{\perp R}$ if and only if c is a solution of the homogeneous system $Ax = 0$.

Proposition 9.72. *If (V, f) is an inner product space, then f is nondegenerate if and only if $V^{\perp L} = \{0\} = V^{\perp R}$; that is, if $f(b, c) = 0$ for all $c \in V$, then $b = 0$, and if $f(b, c) = 0$ for all $b \in V$, then $c = 0$.*

Proof. Our remarks above show that $b \in V^{\perp L}$ if and only if b is a solution of the homogeneous system $A^t x = 0$. Therefore, $V^{\perp L} \neq \{0\}$ if and only if there is a nontrivial solution b, and Exercise 3.70 on page 170 shows that this holds if and only if $\det(A^t) = 0$. Since $\det(A^t) = \det(A)$, we have f degenerate. A similar argument shows that $V^{\perp R} \neq \{0\}$ if and only if there is a nontrivial solution to $Ax = 0$. •

Example 9.73.
Let (V, f) be an inner product space, and let $W \subseteq V$ be a subspace. It is possible that f is nondegenerate, while its restriction $f|(W \times W)$ is degenerate. For example, let $V = k^2$, and let f have the inner product matrix $A = \left[\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right]$ relative to the standard basis e_1, e_2. It is clear that A is nonsingular, so that f is nondegenerate. On the other hand, if $W = \langle e_1 \rangle$, then $f|(W \times W) = 0$, and hence it is degenerate. ◂

Here is a characterization of nondegeneracy in terms of the dual space. This is quite natural, for if f is a bilinear form on a vector space V over a field k, then for any fixed $u \in V$, the function $f(\ , u) \colon V \to k$ is a linear functional.

Proposition 9.74. *Let (V, f) be an inner product space, and let $e_1, \ldots, e_n$ be a basis of V. Then f is nondegenerate if and only if $f(\ , e_1), \ldots, f(\ , e_n)$ is a basis of the dual space V^* (we call the latter the* ***dual basis****).*

Proof. Assume that f is nondegenerate. Since $\dim(V^*) = n$, it suffices to prove linear independence. If there are scalars $c_1, \ldots, c_n$ with $\sum_i c_i f(\ , e_i) = 0$, then

$$\sum_i c_i f(v, e_i) = 0 \quad \text{for all } v \in V.$$

If we define $u = \sum_i c_i e_i$, then $f(v, u) = 0$ for all v, so that nondegeneracy gives $u = 0$. But $e_1, \ldots, e_n$ is a linearly independent list, so that all $c_i = 0$; hence, $f(\ , e_1), \ldots, f(\ , e_n)$ is also linearly independent, and hence it is a basis of V^*.

Conversely, assume the given linear functionals are a basis of V^*. If $f(v, u) = 0$ for all $v \in V$, where $u = \sum_i c_i e_i$, then $\sum_i c_i f(\ , e_i) = 0$. Since these linear functionals are linearly independent, all $c_i = 0$, and so $u = 0$; that is, f is nondegenerate. •

Corollary 9.75. *If (V, f) is an inner product space with f nondegenerate, then every linear functional $g \in V^*$ has the form*

$$g = f(\ , u)$$

for a unique $u \in V$.

Proof. Let $e_1, \ldots, e_n$ be a basis of V, and let $f(\ \ , e_1), \ldots, f(\ \ , e_n)$ be its dual basis. Since $g \in V^*$, there are scalars c_i with $g = \sum_i c_i f(\ \ , e_i)$. If we define $u = \sum_i c_i e_i$, then $g(v) = f(v, u)$.

To prove uniqueness, suppose that $f(\ \ , u) = f(\ \ , u')$. Then $f(v, u - u') = 0$ for all $v \in V$, and so nondegeneracy of f gives $u - u' = 0$. •

Corollary 9.76. *Let (V, f) be an inner product space with f nondegenerate. If $e_1, \ldots, e_n$ is a basis of V, then there exists a basis $b_1, \ldots, b_n$ of V with*

$$f(e_i, b_j) = \delta_{ij}.$$

Proof. Since f is nondegenerate, the function $V \to V^*$, given by $v \mapsto f(\ , v)$, is an isomorphism. It follows that the following diagram commutes:

$$\begin{array}{ccc} V \times V & \xrightarrow{f} & k, \\ \varphi \downarrow & \nearrow_{\text{ev}} & \\ V \times V^* & & \end{array}$$

where ev is evaluation $(x, g) \mapsto g(x)$ and $\varphi\colon (x, y) \mapsto (x, f(\ , y))$. For each i, let $g_i \in V^*$ be the ith coordinate function: If $v \in V$ and $v = \sum_j c_j e_j$, then $g_i(v) = c_i$. By Corollary 9.75, there are $b_1, \ldots, b_n \in V$ with $g_i = f(\ , b_i)$ for all i. Commutativity of the diagram gives

$$f(e_i, b_j) = \text{ev}(e_i, g_j) = \delta_{ij}. \quad \bullet$$

Proposition 9.77. *Let (V, f) be an inner product space, and let W be a subspace of V. If $f|(W \times W)$ is nondegenerate, then*

$$V = W \oplus W^\perp.$$

Remark. We do not assume that f itself is nondegenerate; even if we did, it would not force $f|(W \times W)$ to be nondegenerate, as we have seen in Example 9.73. ◂

Proof. If $u \in W \cap W^\perp$, then $f(w, u) = 0$ for all $w \in W$. Since $f|(W \times W)$ is nondegenerate and $u \in W$, we have $u = 0$; hence, $W \cap W^\perp = \{0\}$. If $v \in V$, then $f(\ \ , v)|W$ is a linear functional on W; that is, $f(\ \ , v)|W \in W^*$. By Corollary 9.75, there is $w_0 \in W$ with $f(w, v) = f(w, w_0)$ for all $w \in W$. Hence, $v = w_0 + (v - w_0)$, where $w_0 \in W$ and $v - w_0 \in W^\perp$. •

There is a name for direct sum decompositions as in the proposition.

Definition. If (V, f) is an inner product space, then we say that a direct sum

$$V = W_1 \oplus \cdots \oplus W_r$$

is an ***orthogonal direct sum*** if, for all $i \neq j$, we have $f(w_i, w_j) = 0$ for all $w_i \in W_i$ and $w_j \in W_j$.

We are now going to look more carefully at special bilinear forms; first we examine alternating forms, then symmetric ones.

We begin by constructing all alternating bilinear forms f on a two-dimensional vector space V over a field k. As always, $f = 0$ is an example. Otherwise, there exist two vectors $e_1, e_2 \in V$ with $f(e_1, e_2) \neq 0$; say, $f(e_1, e_2) = c$. If we replace e_1 by $e_1' = c^{-1}e_1$, then $f(e_1', e_2) = 1$. Since f is alternating, the inner product matrix A of f relative to the basis e_1', e_2 is $A = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$.

Definition. A ***hyperbolic plane*** over a field k is a two-dimensional vector space over k equipped with a nonzero alternating bilinear form.

We have just seen that every two-dimensional alternating space (V, f) in which f is not identically zero has an inner product matrix $A = \left[\begin{smallmatrix} 0 & 1 \\ -1 & 0 \end{smallmatrix}\right]$.

Theorem 9.78. *Let (V, f) be an alternating space, where V is a vector space over a field k. If f is nondegenerate, then there is an orthogonal direct sum*

$$V = H_1 \oplus \cdots \oplus H_m,$$

where each H_i is a hyperbolic plane.

Proof. The proof is by induction on $\dim(V) \geq 1$. For the base step, note that $\dim(V) \geq 2$, because an alternating form on a one-dimensional space must be 0, hence degenerate. If $\dim(V) = 2$, then we saw that V is a hyperbolic plane. For the inductive step, note that there are vectors $e_1, e_2 \in V$ with $f(e_1, e_2) \neq 0$ (because f is nondegenerate, hence nonzero), and we may normalize so that $f(e_1, e_2) = 1$: if $f(e_1, e_2) = d$, replace e_2 by $d^{-1}e_2$. The subspace $H_1 = \langle e_1, e_2 \rangle$ is a hyperbolic plane, and the restriction $f|(H_1 \times H_1)$ is nondegenerate. Thus, Proposition 9.77 gives $V = H_1 \oplus H_1^\perp$. Since the restriction of f to $H_1^\perp$ is nondegenerate, by Exercise 9.56 on page 713, the inductive hypothesis applies. •

Corollary 9.79. *Let (V, f) be an alternating space, where V is a vector space over a field k. If f is nondegenerate, then* $\dim(V)$ *is even.*

Proof. By the theorem, V is a direct sum of two-dimensional subspaces. •

Definition. Let (V, f) be an alternating space in which f is nondegenerate. A ***symplectic***[13] ***basis*** is a basis $x_1, y_1, \ldots, x_m, y_m$ such that $f(x_i, y_i) = 1$, $f(y_i, x_i) = -1$ for all i, and all other $f(x_i, x_j)$, $f(y_i, y_j)$, $f(x_i, y_j)$, and $f(y_j, x_i)$ are 0.

---

[13]The term *symplectic* was coined by H. Weyl. On page 165 of his book, *The Classical Groups; Their Invariants and Representations*, he wrote, "The name 'complex group' formerly advocated by me in allusion to line complexes, as these are defined by the vanishing of antisymmetric bilinear forms, has become more and more embarrassing through collision with the word 'complex' in the connotation of complex number. I therefore propose to replace it by the corresponding Greek adjective 'symplectic.' Dickson calls the group the 'Abelian linear group' in homage to Abel who first studied it."

Corollary 9.80. *Let (V, f) be an alternating space in which f is nondegenerate,*[14] *and let A be an inner product matrix for f (relative to some basis of V).*

(i) *There exists a symplectic basis $x_1, y_1, \ldots, x_m, y_m$ for V, and A is a $2m \times 2m$ matrix for some $m \geq 1$.*

(ii) *A is congruent to a matrix direct sum of blocks of the form $\begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$, and the latter is congruent to $\begin{bmatrix} 0 & I \\ -I & 0 \end{bmatrix}$, where I is the $m \times m$ identity matrix.*

(iii) *Every nonsingular skew-symmetric matrix A over a field k is congruent to a direct sum of 2×2 blocks $\begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$.*

Proof. (i) A symplectic basis exists, by Theorem 9.78, and so V is even-dimensional.

(ii) An inner product matrix A is congruent to the inner product matrix relative to a symplectic basis arising from a symplectic basis $x_1, y_1, \ldots, x_m, y_m$. The second inner product matrix arises from a reordered symplectic basis $x_1, \ldots, x_m, y_1, \ldots, y_m$.

(iii) A routine calculation. •

We now consider symmetric bilinear forms.

Definition. Let (V, f) be a symmetric space, and let $E = e_1, \ldots, e_n$ be a basis of V. Then E is an ***orthogonal basis*** if $f(e_i, e_j) = 0$ for all $i \neq j$, and E is an ***orthonormal basis*** if $f(e_i, e_j) = \delta_{ij}$, where δ_{ij} is the Kronecker delta.

If $e_1, \ldots, e_n$ is an orthogonal basis of a symmetric space (V, f), then $V = \langle e_1 \rangle \oplus \cdots \oplus \langle e_n \rangle$ is an orthogonal direct sum. In Corollary 9.76, we saw that if (V, f) is a symmetric space with f nondegenerate, and if $e_1, \ldots, e_n$ is a basis of V, then there exists a basis $b_1, \ldots, b_n$ of V with $f(e_i, b_j) = \delta_{ij}$. If E is an orthonormal basis, then we can set $b_i = e_i$ for all i.

Theorem 9.81. *Let (V, f) be a symmetric space, where V is a vector space over a field k of characteristic not 2.*

(i) *V has an orthogonal basis, and so every symmetric matrix A with entries in k is congruent to a diagonal matrix.*

(ii) *If $C = \operatorname{diag}[c_1^2 d_1, \ldots, c_n^2 d_n]$, then C is congruent to $D = \operatorname{diag}[d_1, \ldots, d_n]$.*

(iii) *If f is nondegenerate and if every element in k has a square root in k, then V has an orthonormal basis. Every symmetric matrix A with entries in k is congruent to I.*

---

[14]If the form f is degenerate, then A is congruent to a direct sum of 2×2 blocks $\begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$ and a block of 0's.

Proof. (i) If $f = 0$, then every basis is an orthogonal basis. We may now assume that $f \neq 0$. By Exercise 9.54 on page 713, which applies because k does not have characteristic 2, there is some $v \in V$ with $f(v, v) \neq 0$ (otherwise, f is both symmetric and alternating). If $W = \langle v \rangle$, then $f|(W \times W)$ is nondegenerate, so that Proposition 9.77 gives $V = W \oplus W^{\perp}$. The proof is now completed by induction on $\dim(W)$.

If A is an $n \times n$ symmetric matrix, then Proposition 9.70(i) shows that there is a symmetric bilinear form f and a basis $U = u_1, \ldots, u_n$ so that A is the inner product matrix of f relative to U. We have just seen that there exists an orthogonal basis $v_1, \ldots, v_n$, so that Proposition 9.70(iii) shows that A is congruent to the diagonal matrix $\operatorname{diag}[f(v_1, v_1), \ldots, f(v_n, v_n)]$.

(ii) If an orthogonal basis consists of vectors v_i with $f(v_i, v_i) = c_i^2 d_i$, then replacing each v_i by $v_i' = c_i^{-1} v_i$ gives an orthogonal basis with $f(v_i', v_i') = d_i$. It follows that the inner product matrix of f relative to the basis $v_1', \ldots, v_n'$ is $D = \operatorname{diag}[d_1, \ldots, d_n]$.

(iii) By part (i), there exists an orthogonal basis $v_1, \ldots, v_n$; let $f(v_i, v_i) = c_i$ for each i. Since f is nondegenerate, $c_i \neq 0$ for all i (the determinant of the inner product matrix relative to this orthogonal basis is $c_1 c_2 \cdots c_n$); since each c_i is a square, by hypothesis, we may replace each v_i by $v_i' = (\sqrt{c_i})^{-1} v_i$, as in part (ii); this new basis is orthonormal. The final statement follows because the inner product matrix relative to an orthonormal basis is the identity I. •

Notice that Theorem 9.81 does not say that any two diagonal matrices over a field k of characteristic not 2 are congruent; this depends on k. For example, if $k = \mathbb{C}$, then all (nonsingular) diagonal matrices are congruent to I, but we now show that this is false if $k = \mathbb{R}$.

Definition. A symmetric bilinear form f on a vector space V over $\mathbb{R}$ is ***positive definite*** if $f(v, v) > 0$ for all nonzero $v \in V$, while f is ***negative definite*** if $f(v, v) < 0$ for all nonzero $v \in V$.

The next result, and its matrix corollary, was proved by J. J. Sylvester. When $n = 2$, it classifies the conic sections, and when $n = 3$, it classifies the quadric surfaces.

Lemma 9.82. *If f is a symmetric bilinear form on a vector space V over $\mathbb{R}$ of dimension m, then there is an orthogonal direct sum*

$$V = W_+ \oplus W_- \oplus W_0,$$

where $f|W_+$ is positive definite, $f|W_-$ is negative definite, and $f|W_0$ is identically 0. *Moreover, the dimensions of these three subspaces are uniquely determined by f.*

Proof. By Theorem 9.81, there is an orthogonal basis $v_1, \ldots, v_m$ of V. Denote $f(v_i, v_i)$ by d_i. As any real number, each d_i is either positive, negative, or 0, and we rearrange the basis vectors so that $v_1, \ldots, v_p$ have postive d_i, $v_{p+1}, \ldots, v_{p+r}$ have negative d_i, and the last vectors have $d_i = 0$. It follows easily that V is the orthogonal direct sum

$$V = \langle v_1, \ldots, v_p \rangle \oplus \langle v_{p+1}, \ldots, v_{p+r} \rangle \oplus \langle v_{p+r+1}, \ldots, v_m \rangle,$$

and that the restrictions of f to each summand are positive definite, negative definite, and zero.

Now $W_0 = V^\perp$ depends only on f, and hence its dimension depends only on f as well. To prove uniqueness of the other two dimensions, suppose that there is a second orthogonal direct sum $V = W'_+ \oplus W'_- \oplus W_0$. If $T : V \to W_+$ is the projection, then $\ker T = W_- \oplus W_0$. It follows that if $\varphi = T|W'_+$, then

$$\ker \varphi = W'_+ \cap \ker T = W'_+ \cap \big(W_- \oplus W_0\big).$$

However, if $v \in W'_+$, then $f(v, v) \geq 0$, while if $v \in W_- \oplus W_0$, then $f(v, v) \leq 0$; hence, if $v \in \ker\varphi$, then $f(v, v) = 0$. But $f|W'_+$ is positive definite, for this is one of the defining properties of W'_+, so that $f(v, v) = 0$ implies $v = 0$. We conclude that $\ker\varphi = \{0\}$, and so $\varphi\colon W'_+ \to W_+$ is an injection; therefore, $\dim(W'_+) \leq \dim(W_+)$. The reverse inequality is proved similarly, so that $\dim(W'_+) = \dim(W_+)$. Finally, the formula $\dim(W_-) = \dim(V) - \dim(W_+) - \dim(W_0)$, and its primed version, give $\dim(W'_-) = \dim(W_-)$. •

Theorem 9.83 (Law of Inertia). *Every symmetric $n \times n$ matrix A over $\mathbb{R}$ is congruent to a matrix of the form*

$$\begin{bmatrix} I_p & 0 & 0 \\ 0 & -I_r & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

Moreover, the ***signature*** *s of f, defined by $s = p - r$ is well-defined, and two $n \times n$ symmetric real matrices are congruent if and only if they have the same rank and the same signature.*

Proof. By Theorem 9.81, A is congruent to a diagonal matrix $\text{diag}[d_1, \ldots, d_n]$, where $d_1, \ldots, d_p$ are positive, $d_{p+1}, \ldots, d_{p+r}$ are negative, and $d_{p+r+1}, \ldots, d_n$ are 0. But every positive real is a square, while every negative real is the negative of a square; it now follows from Theorem 9.81(ii) that A is congruent to a matrix as in the statement of the theorem.

It is clear that congruent $n \times n$ matrices have the same rank and the same signature. Conversely, let A and A' have the same rank and the same signature. Now A is congruent to the matrix direct sum $I_p \oplus -I_r \oplus 0$ and A' is congruent to $I_{p'} \oplus -I_{r'} \oplus 0$. Since $\text{rank}(A) = \text{rank}(A')$, we have $p' + r' = p + r$; since the signatures are the same, we have $p' - r' = p - r$. It follows that $p' = p$ and $r' = r$, so that both A and A' are congruent to the same diagonal matrix of 1's, -1's, and 0's, and hence they are congruent to each other. •

It would be simplest if a symmetric space (V, f) with f nondegenerate always had an orthonormal basis; that is, if every symmmetric matrix were congruent to the identity matrix. This need not be so, for the 2×2 real matrix $-I$ is not congruent to I because their signatures are different (I has signature 2 and $-I$ has signature -2).

Closely related to bilinear forms are *quadratic forms*; they arise from a bilinear form f defined on a vector space V over a field k by considering the function $Q\colon V \to k$ given by $Q(v) = f(v, v)$.

Definition. Let V be an a vector space over a field k. A ***quadratic form*** is a function $Q\colon V \to k$ such that

(i) $Q(cv) = c^2 Q(v)$ for all $v \in V$ and $c \in k$;

(ii) the function $g\colon V \times V \to k$, defined by

$$g(u, v) = Q(u + v) - Q(u) - Q(v),$$

is a bilinear form.

Example 9.84.
(i) If f is a bilinear form on a vector space V, define $Q(v) = f(v, v)$ for all $v \in V$; we show that Q is a quadratic form. Now $Q(cv) = f(cv, cv) = c^2 f(v, v) = c^2 Q(v)$, giving the first axiom in the definition. If $u, v \in V$, then

$$\begin{aligned} Q(u + v) &= f(u + v, u + v) \\ &= f(u, u) + f(u, v) + f(v, u) + f(v, v) \\ &= Q(u) + Q(v) + g(u, v), \end{aligned}$$

where

$$g(u, v) = f(u, v) + f(v, u).$$

It is easy to check that g is a bilinear form that is symmetric.

(ii) We have just seen that every bilinear form f determines a quadratic form Q. If f is symmetric and k does not have characteristic 2, then Q determines f. In fact, the formula $2f(u, v) = g(u, v)$ gives $f(u, v) = \frac{1}{2} g(u, v)$ in this case.

(iii) If f is the usual dot product defined on $\mathbb{R}^n$, then the corresponding quadratic form is $Q(v) = \|v\|^2$, where $\|v\|$ is the length of the vector v.

(iv) If f is a bilinear form on a vector space V with inner product matrix $A = [a_{ij}]$ relative to some basis $e_1, \ldots, e_n$, then if $u = \sum c_i e_i$,

$$Q(u) = \sum_{i,j} a_{ij} c_i c_j.$$

If $n = 2$, for example, we have

$$Q(u) = a_{11} c_1^2 + (a_{12} + a_{21}) c_1 c_2 + a_{22} c_2^2.$$

Thus, quadratic forms are really homogeneous quadratic polynomials. ◂

We have just observed, in the last example, that if a field k does not have characteristic 2, then symmetric bilinear forms and quadratic forms are merely two different ways of viewing the same thing, for each determines the other.

We have classified quadratic forms Q over $\mathbb{C}$ and over $\mathbb{R}$. The classification over the prime fields is also known, as is the classification over the finite fields, and we now state (without proof) the results when Q is nondegenerate. Given a quadratic form Q defined on a finite-dimensional vector space V over a field k, its associated bilinear form is

$$f(x, y) = Q(x + y) - Q(x) - Q(y).$$

Call two quadratic forms ***equivalent*** if their associated bilinear forms have congruent inner product matrices, and call a quadratic form ***nondegenerate*** if its bilinear form is nondegerate. As we have just seen in Example 9.84, f is a symmetric bilinear form (which is uniquely determined by Q when k does not have characteristic 2). If k is a finite field of odd characteristic, then two nondegenerate quadratic forms over k are equivalent if and only if they have the same discriminant (see Kaplansky, *Linear Algebra and Geometry; A Second Course*, pp. 14–15). If k is a finite field of characteristic 2, the theory is a bit more complicated. In this case, the associated symmetric bilinear form

$$g(x, y) = Q(x + y) + Q(x) + Q(y)$$

must also be alternating, for $g(x, x) = Q(2x)+2Q(x) = 0$. Therefore, V has a symplectic basis $x_1, y_1, \dots, x_m, y_m$. The ***Arf invariant*** of Q is defined by

$$\mathrm{Arf}(Q) = \sum_{i=1}^{m} Q(x_i)Q(y_i)$$

[it is not at all obvious that the Arf invariant is an invariant; i.e., that $\mathrm{Arf}(Q)$ does not depend on the choice of symplectic basis; see R. L. Dye, “On the Arf Invariant,” *Journal of Algebra* 53 (1978), pp. 36–39, for an elegant proof]. If k is a finite field of characteristic 2, then two nondegenerate quadratic forms over k are equivalent if and only if they have the same discriminant and the same Arf invariant (see Kaplansky, *Linear Algebra and Geometry; A Second Course*, pp. 27–33). The classification of quadratic forms over $\mathbb{Q}$ is much deeper (see Borevich and Shafarevich, *Number Theory*, pp. 61–70). Just as $\mathbb{R}$ can be obtained from $\mathbb{Q}$ by completing with respect to the usual metric $d(a, b) = |a - b|$ (that is, by adding points to force Cauchy sequences to converge), so, too, can we complete $\mathbb{Z}$, for every prime p, with respect to the p-adic metric (see the discussion on page 503). The completion $\mathbb{Z}_p$ is called the ***p-adic integers***. The p-adic metric on $\mathbb{Z}$ can be extended to $\mathbb{Q}$, and its completion $\mathbb{Q}_p$ [which turns out to be $\mathrm{Frac}(\mathbb{Z}_p)$] is called the ***p-adic numbers***. The ***Hasse–Minkowski theorem*** says that two quadratic forms over $\mathbb{Q}$ are equivalent if and only if they are equivalent over $\mathbb{R}$ and over $\mathbb{Q}_p$ for all primes p.

The first theorems of linear algebra consider the structure of vector spaces in order to pave the way for a discussion of linear transformations. Similarly, the first theorems of inner product spaces enable us to discuss the appropriate linear transformations.

Definition. If (V, f) is an inner product space, where V is a finite-dimensional vector space over a field k and f is a nondegenerate bilinear form, then an ***isometry*** is a linear transformation $\varphi\colon V \to V$ such that, for all $u, v \in V$,

$$f(u, v) = f(\varphi u, \varphi v).$$

Proposition 9.85. *Let (V, f) be an inner product space, where f is a nondegenerate bilinear form, let $E = e_1, \ldots, e_n$ be a basis of V, and let A be the inner product matrix relative to E. Then $\varphi \in \mathrm{GL}(V)$ is an isometry if and only if its matrix $M = {}_E[\varphi]_E$ satisfies the equation $M^t AM = A$.*

Proof. Recall the equation

$$f(b, c) = B^t AC,$$

where $b, c \in V$ and $B, C \in k^n$ are their coordinate vectors relative to the basis E. In this notation, $E_1, \ldots, E_n$ is the standard basis of k^n. Now

$$\varphi(e_i) = ME_i$$

for all i, because ME_i is the ith column of M that is the coordinate vector of $\varphi(e_i)$. Therefore,

$$f(\varphi e_i, \varphi e_j) = (ME_i)^t A(ME_j) = E_i^t(M^t AM)E_j.$$

If φ is an isometry, then

$$f(\varphi e_i, \varphi e_j) = f(e_i, e_j) = E_i^t AE_j,$$

so that $f(e_i, e_j) = E_i^t AE_j = E_i^t(M^t AM)E_j$. Hence, Proposition 9.70(ii) gives $M^t AM = A$.

Conversely, if $M^t AM = A$, then

$$f(\varphi e_i, \varphi e_j) = E_i^t(M^t AM)E_j = E_i^t AE_j = f(e_i, e_j),$$

and φ is an isometry. •

Proposition 9.86. *Let (V, f) be an inner product space, where V is a finite-dimensional vector space over a field k and f is a nondegenerate bilinear form. Then $\mathrm{Isom}(V, f)$, the set of all isometries of (V, f), is a subgroup of $\mathrm{GL}(V)$.*

Proof. We prove that $\mathrm{Isom}(V, f)$ is a subgroup; of course, 1_V is an isometry. Let $\varphi \colon V \to V$ be an isometry. If $u \in V$ and $\varphi u = 0$, then, for all $v \in V$, we have

$$0 = f(\varphi u, \varphi v) = f(u, v).$$

Since f is nondegenerate, $u = 0$ and φ is an injection. Hence, $\dim(\mathrm{im}\,\varphi) = \dim(V)$, so that $\mathrm{im}\,\varphi = V$, by Corollary 3.90(ii). Therefore, every isometry is nonsingular.

The inverse of an isometry φ is also an isometry: For all $u, v \in V$,

$$\begin{aligned} f(\varphi^{-1}u, \varphi^{-1}v) &= f(\varphi\varphi^{-1}u, \varphi\varphi^{-1}v) \\ &= f(u, v). \end{aligned}$$

Finally, the composite of two isometries φ and θ is also an isometry:

$$f(u, v) = f(\varphi u, \varphi v) = f(\theta\varphi u, \theta\varphi v). \quad \bullet$$

Computing the inverse of a general nonsingular matrix is quite time-consuming, but it is easier for isometries.

Definition. Let (V, f) be an inner product space whose bilinear form f is nondegenerate. The ***adjoint*** of a linear transformation $T : V \to V$ is a linear transformation $T^* : V \to V$ such that, for all $u, v \in V$,

$$f(Tu, v) = f(u, T^*v).$$

Let us see that adjoints exist.

Proposition 9.87. *If (V, f) is an inner product space whose bilinear form f is nondegenerate, then every linear transformation $T : V \to V$ has an adjoint.*

Proof. Let $e_1, \ldots, e_n$ be a basis of V. For each j, the function $\varphi_j : V \to k$, defined by

$$\varphi_j(v) = f(Tv, e_j),$$

is easily seen to be a linear functional. By Corollary 9.75, there exists $u_j \in V$ with $\varphi_j(v) = f(v, u_j)$ for all $v \in V$. Define $T^* : V \to V$ by $T^*(e_j) = u_j$, and note that

$$f(Te_i, e_j) = \varphi_j(e_i) = f(e_i, u_j) = f(e_i, T^*e_j). \quad \bullet$$

Proposition 9.88. *Let (V, f) be an inner product space whose bilinear form f is nondegenerate. If $T : V \to V$ is a linear transformation with adjoint T^*, then T is an isometry if and only if $T^*T = 1_V$, in which case $T^* = T^{-1}$.*

Proof. If $T^*T = 1_V$, then, for all $u, v \in V$, we have

$$f(Tu, Tv) = f(u, T^*Tv) = f(u, v),$$

so that T is an isometry.

Conversely, assume that T is an isometry. Choose $v \in V$; for all $u \in V$, we have

$$\begin{aligned} f(u, T^*Tv - v) &= f(u, T^*Tv) - f(u, v) \\ &= f(Tu, Tv) - f(u, v) \\ &= 0. \end{aligned}$$

Since f is nondegenerate, $T^*Tv - v = 0$; that is, $T^*Tv = v$. As this is true for all $v \in V$, we have $T^*T = 1_V$. $\bullet$

Definition. Let (V, f) be an inner product space, where V is a finite-dimensional vector space over a field k and f is a nondegenerate bilinear form.

(i) If f is alternating, then $\text{Isom}(V, f)$ is called the ***symplectic group***, and it is denoted by $\text{Sp}(V, f)$.

(ii) If f is symmetric, then $\text{Isom}(V, f)$ is called the ***orthogonal group***, and it is denoted by $O(V, f)$.

As always, a choice of basis E of an n-dimensional vector space V over a field k gives an isomorphism $\mu\colon \mathrm{GL}(V) \to \mathrm{GL}(n, k)$, the group of all nonsingular $n \times n$ matrices over k. In particular, let (V, f) be an alternating space with f nondegenerate, and let $E = x_1, y_1, \ldots, x_m, y_m$ be a symplectic basis of V (which exists, by Corollary 9.80); recall that $n = \dim(V)$ is even; say, $n = 2m$. Denote the image of $\mathrm{Sp}(V, f)$ by $\mathrm{Sp}(2m, k)$. Similarly, if (V, f) is a symmetric space with f nondegenerate, and E is an orthogonal basis (which exists when k does not have characteristic 2, by Theorem 9.81), denote the image of $O(V, f)$ by $O(n, f)$.

Let us find adjoints when the bilinear form is symmetric or alternating.

Proposition 9.89. *Let (V, f) be a symmetric space, where V is an n-dimensional vector space over a field k and f is nondegenerate, and let $E = e_1, \ldots, e_n$ be an orthogonal basis with $f(e_i, e_i) = c_i$.*

If $B = [b_{ij}]$ is a matrix relative to E, then its adjoint B^* is its "weighted" transpose $[c_i^{-1} c_j b_{ji}]$. In particular, if E is an orthonormal basis, then $B^* = B^t$, the transpose of B.

Remark. It follows that B is orthogonal if and only if $B^t B = I$. ◀

Proof. We have

$$\begin{aligned} f(Be_i, e_j) &= f\Big(\sum_\ell b_{\ell i} e_\ell, e_j\Big) \\ &= \sum_\ell b_{\ell i} f(e_\ell, e_j) \\ &= b_{ji} c_j. \end{aligned}$$

If $B^* = [b^*_{ij}]$, then a similar calculation gives

$$f(e_i, B^* e_j) = \sum b^*_{\ell j} f(e_i, e_\ell) = c_i b^*_{ij}.$$

Since $f(Be_i, e_j) = f(e_i, B^* e_j)$, we have

$$b_{ji} c_j = c_i b^*_{ij}$$

for all i, j. Since f is nondegenerate, all $c_i \neq 0$, and so

$$b^*_{ij} = c_i^{-1} c_j b_{ji}.$$

The last remark follows, for if E is an orthonormal basis, then $c_i = 1$ for all i. ●

How can we recognize a symplectic matrix?

Proposition 9.90. *Let (V, f) be an alternating space, where V is a $2m$-dimensional vector space over a field k and f is nondegenerate, and let E be a symplectic basis ordered as $x_1, \ldots, x_m, y_1, \ldots, y_m$.*

The adjoint of a matrix $B = \begin{bmatrix} P & Q \\ S & T \end{bmatrix}$ relative to E, partitioned into $m \times m$ blocks, is

$$B^* = \begin{bmatrix} T^t & -Q^t \\ -S^t & P^t \end{bmatrix}.$$

Remark. It follows that $B \in \mathrm{Sp}(2m, k)$ if and only if $B^*B = I$. ◂

Proof. We have

$$\begin{aligned} f(Bx_i, x_j) &= f\Big(\sum_\ell p_{\ell i} x_\ell + s_{\ell i} y_\ell, x_j\Big) \\ &= \sum_\ell p_{\ell i}\, f(x_\ell, x_j) + \sum_\ell s_{\ell i}\, f(y_\ell, x_j) \\ &= -s_{ji}, \end{aligned}$$

because $f(x_\ell, x_j) = 0$ and $f(y_\ell, x_j) = -\delta_{\ell j}$ for all i, j. Let us partition the adjoint B^* into $m \times m$ blocks:

$$B^* = \begin{bmatrix} \Pi & K \\ \Sigma & \Omega \end{bmatrix}.$$

Hence,

$$\begin{aligned} f(x_i, B^*x_j) &= f\big(x_i, \sum_\ell \pi_{\ell j} x_\ell + \sigma_{\ell j} y_\ell\big) \\ &= \sum_\ell \pi_{\ell j}\, f(x_i, x_\ell) + \sum_\ell \sigma_{\ell j}\, f(x_i, y_\ell) \\ &= \sigma_{ij}, \end{aligned}$$

because $f(x_i, x_\ell) = 0$ and $f(x_i, y_\ell) = \delta_{i\ell}$. Since $f(Bx_i, x_j) = f(x_i, B^*x_j)$, we have $\sigma_{ij} = -s_{ji}$. Hence, $\Sigma = -S^t$. Computation of the other blocks of B^* is similar. •

The next question is whether $\mathrm{Isom}(V, f)$ depends on the choice of nondegenerate alternating bilinear form f. Observe that $\mathrm{GL}(V)$ acts on $k^{V \times V}$, the set of all functions $V \times V \to k$: If $f \colon V \times V \to k$ and $\varphi \in \mathrm{GL}(V)$, then define $\varphi f = f^\varphi$, where

$$f^\varphi(b, c) = f(\varphi^{-1}b, \varphi^{-1}c).$$

This formula does yield an action: If $\theta \in \mathrm{GL}(V)$, then $(\varphi\theta) f = f^{\varphi\theta}$, where

$$\begin{aligned} (\varphi\theta) f(b, c) &= f^{\varphi\theta}(b, c) \\ &= f((\varphi\theta)^{-1}b, (\varphi\theta)^{-1}c) \\ &= f(\theta^{-1}\varphi^{-1}b, \theta^{-1}\varphi^{-1}c). \end{aligned}$$

On the other hand, $\varphi(\theta f)$ is defined by

$$\begin{aligned}(f^\theta)^\varphi(b,c) &= f^\theta(\varphi^{-1}b, \varphi^{-1}c)\\ &= f(\theta^{-1}\varphi^{-1}b, \theta^{-1}\varphi^{-1}c),\end{aligned}$$

so that $(\varphi\theta)f = \varphi(\theta f)$.

Definition. Let V and W be finite-dimensional vector spaces over a field k, and let $f: V \times V \to k$ and $g: W \times W \to k$ be bilinear forms. Then f and g are ***equivalent*** if there is an isometry $\varphi: V \to W$.

Proposition 9.91. *If V is a finite-dimensional vector space over a field k and if $f, g: V \times V \to k$ are bilinear forms, then the following statements are equivalent.*

(i) *f and g are equivalent.*

(ii) *If $E = e_1, \ldots, e_n$ is a basis of V, then the inner product matrices of f and g with respect to E are congruent.*

(iii) *There is $\varphi \in \mathrm{GL}(V)$ with $g = f^\varphi$.*

Proof. (i) $\Rightarrow$ (ii) If $\varphi: V \to V$ is an isometry, then $g(\varphi(b), \varphi(c)) = f(b,c)$ for all $b, c \in V$. If $E = e_1, \ldots, e_n$ is a basis of V, then $E' = \varphi(e_1), \ldots, \varphi(e_n)$ is also a basis, because φ is an isomorphism. Hence, $A' = [g(\varphi(e_i), \varphi(e_j))] = [f(e_i, e_j)] = A$ for all i, j; that is, the inner product matrix A' of g with respect to E' is equal to the inner product matrix A of f with respect to E. By Proposition 9.70(iii), the inner product matrix A'' of g with respect to E is congruent to A.

(ii) $\Rightarrow$ (iii) If $A = [f(e_i, e_j)]$ and $A' = [g(e_i, e_j)]$, then there exists a nonsingular matrix $Q = [q_{ij}]$ with $A' = Q^t AQ$. Define $\theta: V \to V$ to be the linear transformation with $\theta(e_j) = \sum_\nu q_{\nu j} e_\nu$. Finally, $g = f^{\theta^{-1}}$:

$$\begin{aligned}[g(e_i, e_j)] = A' = Q^t AQ &= \Big[f\Big(\sum_\nu q_{\nu i} e_\nu, \sum_\lambda q_{\lambda j} e_\lambda\Big)\Big]\\ &= [f(\theta(e_i), \theta(e_j))] = [f^{\theta^{-1}}(e_i, e_j)].\end{aligned}$$

(iii) $\Rightarrow$ (i) It is obvious from the definition that $\varphi^{-1}: (V, g) \to (V, f)$ is an isometry:

$$g(b,c) = f^\varphi(b,c) = f(\varphi^{-1}b, \varphi^{-1}c).$$

Therefore, g is equivalent to f. •

Proposition 9.92.

(i) *Let (V, f) be an inner product space, where V is a finite-dimensional vector space over a field k and f is a nondegenerate bilinear form. The stabilizer $\mathrm{GL}(V)_f$ of f under the action on $k^{V \times V}$ is $\mathrm{Isom}(V, f)$.*

(ii) *If $g: V \times V \to k$ lies in the same orbit as f, then $\mathrm{Isom}(V, f)$ and $\mathrm{Isom}(V, g)$ are isomorphic; in fact, they are conjugate subgroups of $\mathrm{GL}(V)$.*

Proof. (i) By definition of stabilizer, $\varphi \in \mathrm{GL}(V)_f$ if and only if $f^\varphi = f$; that is, for all $b, c \in V$, we have $f(\varphi^{-1}b, \varphi^{-1}c) = f(b, c)$. Thus, φ^{-1}, and hence φ, is an isometry.

(ii) By Exercise 2.99 on page 114, we have $\mathrm{GL}(V)_g = \tau(\mathrm{GL}(V)_f)\tau^{-1}$ for some $\tau \in \mathrm{GL}(V)$; that is, $\mathrm{Isom}(V, g) = \tau \mathrm{Isom}(V, f)\tau^{-1}$. •

It follows from Proposition 9.92 that equivalent bilinear forms have isomorphic isometry groups. We can now show that the symplectic group is, to isomorphism, independent of the choice of nondegenerate alternating form.

Theorem 9.93. *If (V, f) and (V, g) are alternating spaces, where f and g are nondegenerate, then f and g are equivalent and*

$$\mathrm{Sp}(V, f) \cong \mathrm{Sp}(V, g).$$

Proof. By Corollary 9.80(ii), the inner product matrix of any nondegenerate alternating bilinear form is congruent to $\left[\begin{smallmatrix} 0 & I \\ -I & 0 \end{smallmatrix}\right]$, where I is the identity matrix. The result now follows from Proposition 9.91. •

Symplectic groups give rise to simple groups. If k is a field, define $\mathrm{PSp}(2m, k) = \mathrm{Sp}(2m, k)/Z(2m, k)$, where $Z(2m, k)$ is the subgroup of all scalar matrices in $\mathrm{Sp}(2m, k)$. The groups $\mathrm{PSp}(2m, k)$ are simple for all $m \geq 1$ and all fields k with only three exceptions: $\mathrm{PSp}(2, \mathbb{F}_2) \cong S_3$, $\mathrm{PSp}(2, \mathbb{F}_3) \cong A_4$, and $\mathrm{PSp}(4, \mathbb{F}_2) \cong S_6$.

The orthogonal groups, that is, isometry groups of a symmetric space (V, f) when f is nondegenerate, also give rise to simple groups. In contrast to symplectic groups, however, they depend on properties of the field k. We restrict our attention to finite fields k. The cases when k has odd characteristic and when k has characteristic 2 must be considered separately, and we must further consider the subcases when $\dim(V)$ is odd or even. When k has odd characteristic p, there is only one orthogonal group $O(n, p^m)$[15] when n is odd, but there are two, $O^+(n, p^m)$ and $O^-(n, p^m)$, when n is even. The simple groups are defined from these groups as follows: First form $SO^\epsilon(n, p^m)$ (where $\epsilon = +$ or $\epsilon = -$) as all orthogonal matrices having determinant 1; next, form $PSO^\epsilon(n, p^m)$ by dividing by all scalar matrices in $SO^\epsilon(n, p^m)$. Finally, we can define a subgroup $\Omega^\epsilon(n, p^m)$ of $PSO^\epsilon(n, p^m)$ (essentially the commutator subgroup), and these groups are simple with only a finite number of exceptions (which can be explicitly listed).

When k has characteristic 2, we usually begin with a quadratic form rather than a symmetric bilinear form. In this case, there is also only one orthogonal group $O(n, 2^m)$ when n is odd, but there are two, which are also denoted by $O^+(n, 2^m)$ and $O^-(n, 2^m)$, when n is even. If n is odd, say, $n = 2\ell + 1$, then $O(2\ell + 1, 2^m) \cong \mathrm{Sp}(2\ell, 2^m)$, so that we consider only orthogonal groups $O^\epsilon(2\ell, 2^m)$ arising from symmetric spaces of even dimension. Each of these groups gives rise to a simple group in a manner analogous to the odd characteristic case. For more details, we refer the reader to the books of E. Artin, *Geometric Algebra*; Conway et al, *Atlas of Finite Groups*; J; Dieudonné, *La Géometrie des groupes*

[15]When k is a finite field, say, $k = \mathbb{F}_q$ for some prime power q, we often denote $\mathrm{GL}(n, k)$ by $\mathrm{GL}(n, q)$. A similar notational change is used for any of the matrix groups arising from $\mathrm{GL}(n, k)$.

classiques; M. Suzuki, *Group Theory* I; and the article by Carter in Kostrikin–Shafarevich, *Algebra* IX.

Quadratic forms are of great importance in number theory. For an introduction to this aspect of the subject, see Hahn, *Quadratic Algebras, Clifford Algebras, and Arithmetic Witt Groups*; Lam, *The Algebraic Theory of Quadratic Forms*; and O'Meara, *Introduction to Quadratic Forms*.

EXERCISES

9.51 It is shown in analytic geometry that if ℓ_1 and ℓ_2 are lines with slopes m_1 and m_2, respectively, then ℓ_1 and ℓ_2 are perpendicular if and only if $m_1 m_2 = -1$. If

$$\ell_i = \{\alpha v_i + u_i : \alpha \in \mathbb{R}\},$$

for $i = 1, 2$, prove that $m_1 m_2 = -1$ if and only if the dot product $v_1 \cdot v_2 = 0$. (Since both lines have slopes, neither of them is vertical.)
Hint. The slope of a vector $v = (a, b)$ is $m = b/a$.

9.52 (i) In calculus, a line in space passing through a point u is defined as

$$\{u + \alpha w : \alpha \in \mathbb{R}\} \subseteq \mathbb{R}^3,$$

where w is a fixed nonzero vector. Show that every line through the origin is a one-dimensional subspace of $\mathbb{R}^3$.

(ii) In calculus, a plane in space passing through a point u is defined as the subset

$$\{v \in \mathbb{R}^3 : (v - u) \cdot n = 0\} \subseteq \mathbb{R}^3,$$

where $n \neq 0$ is a fixed *normal vector*. Prove that a plane through the origin is a two-dimensional subspace of $\mathbb{R}^3$.
Hint. To determine the dimension of a plane through the origin, find an orthogonal basis of $\mathbb{R}^3$ containing n.

9.53 If k is a field of characteristic not 2, prove that for every $n \times n$ matrix A with entries in k, there are unique matrices B and C with B symmetric, C skew-symmetric (i.e., $C^t = -C$), and $A = B + C$.

9.54 Let (V, f) be an inner product space, where V is a vector space over a field k of characteristic not 2. Prove that if f is both symmetric and alternating, then $f = 0$.

9.55 If (V, f) is an inner product space, define $u \perp v$ to mean $f(u, v) = 0$. Prove that $\perp$ is a symmetric relation if and only if f is either symmetric or alternating.

9.56 Let (V, f) be an inner product space with f nondegenerate. If W is a proper subspace and $V = W \oplus W^\perp$, prove that $f|(W^\perp \times W^\perp)$ is nondegenerate.

9.57 (i) Let (V, f) be an inner product space, where V is a vector space over a field k of characteristic not 2. Prove that if f is symmetric, then there is a basis $e_1, \ldots, e_n$ of V and scalars $c_1, \ldots, c_n$ such that $f(x, y) = \sum_i c_i x_i y_i$, where $x = \sum x_i e_i$ and $y = \sum y_i e_i$. Moreover, if f is nondegenerate and k has square roots, then the basis $e_1, \ldots, e_n$ can be chosen so that $f(x, y) = \sum_i x_i y_i$.

(ii) If k is a field of characteristic not 2, then every symmetric matrix A with entries in k is congruent to a diagonal matrix. Moreover, if A is nonsingular and k has square roots, then $A = P^t P$ for some nonsingular matrix P.

9.58 Give an example of two real symmetric $m \times m$ matrices having the same rank and the same discriminant but that are not congruent.

9.59 For every field k, prove that $\mathrm{Sp}(2, k) = \mathrm{SL}(2, k)$.
Hint. By Corollary 9.80(ii), we know that if $P \in \mathrm{Sp}(2m, k)$, then $\det(P) = \pm 1$. However, Proposition 9.89 shows that $\det(P) = 1$ for $P \in \mathrm{Sp}(2, k)$ [it is true, for all $m \geq 1$, that $\mathrm{Sp}(2m, k) \leq \mathrm{SL}(2m, k)$].

9.60 If A is an $m \times m$ matrix with $A^t A = I$, prove that $\begin{bmatrix} A & 0 \\ 0 & A \end{bmatrix}$ is a symplectic matrix. Conclude, if k is a finite field of odd characteristic, that $O(m, k) \leq \mathrm{Sp}(2m, k)$.

9.61 Let (V, f) be an alternating space with f nondegenerate. Prove that $T \in \mathrm{GL}(V)$ is an isometry [i.e., $T \in \mathrm{Sp}(V, f)$] if and only if, whenever $E = x_1, y_1, \ldots, x_m, y_m$ is a symplectic basis of V, then $T(E) = Tx_1, Ty_1, \ldots, Tx_m, Ty_m$ is also a symplectic basis of V.

9.6 Graded Algebras

We are now going to use tensor products of many modules in order to construct some useful rings. This topic is often called *multilinear algebra*.

Throughout this section, R will denote a commutative ring.

Definition. An R-algebra A is a ***graded R-algebra*** if there are R-submodules A^p, for $p \geq 0$, such that

(i) $A = \sum_{p \geq 0} A^p$;

(ii) For all $p, q \geq 0$, if $x \in A^p$ and $y \in A^q$, then $xy \in A^{p+q}$; that is,

$$A^p A^q \subseteq A^{p+q}.$$

An element $x \in A^p$ is called ***homogeneous*** of ***degree*** p.

Notice that 0 is homogeneous of any degree, but that most elements in a graded ring are not homogeneous and, hence, have no degree. Note also that any product of homogeneous elements is itself homogeneous.

Example 9.94.
(i) The polynomial ring $A = R[x]$ is a graded R-algebra if we define

$$A^p = \{rx^p : r \in R\}.$$

The homogeneous elements are the monomials and, in contrast to ordinary usage, only monomials (including 0) have degrees. On the other hand, x^p has degree p in both usages of the term *degree*.

(ii) The polynomial ring $A = R[x_1, x_2, \ldots, x_n]$ is a graded R-algebra if we define

$$A^p = \left\{ r x_1^{e_1} x_2^{e_2} \cdots x_n^{e_n} : r \in R \text{ and } \sum e_i = p \right\};$$

that is, A^p consists of all monomials of total degree p.

(iii) In algebraic topology, we assign a sequence of (abelian) *cohomology groups* $H^p(X, R)$ to a space X, where R is a commutative ring and $p \geq 0$, and we define a multiplication on $\sum_{p \geq 0} H^p(X, R)$, called *cup product*, making it a graded R-algebra. ◂

Just as the degree of a polynomial is often useful, so, too, is the degree of a homogeneous element in a graded algebra.

Definition. If A and B are graded R-algebras, then a ***graded map***[16] is an R-algebra map $f : A \to B$ with $f(A^p) \subseteq B^p$ for all $p \geq 0$.

It is easy to see that all graded R-algebras and graded maps form a category, which we denote by $\mathbf{Gr}_R\mathbf{Alg}$.

Definition. If A is a graded R-algebra, then a ***graded ideal*** (or ***homogeneous ideal***) is a two-sided ideal I in A with $I = \sum_{p \geq 0} I^p$, where $I^p = I \cap A^p$.

In contrast to the *affine varieties* that we have considered in Chapter 6, ***projective varieties*** are studied more intensely in algebraic geometry. The algebraic way to study these geometric objects involves homogeneous ideals in graded algebras.

Proposition 9.95. *Let A and B be graded R-algebras.*

(i) *If $f : A \to B$ is a graded map, then* $\ker f$ *is a graded ideal.*

(ii) *If I is a graded ideal in A, then A/I is a graded R-algebra if we define*

$$(A/I)^p = (A^p + I)/I.$$

Moreover, $A/I = \sum_p (A/I)^p \cong \sum_p A^p/(I \cap A^p) = \sum_p (A^p/I^p)$.

(iii) *A two-sided ideal I in A is graded if and only if it is generated by homogeneous elements.*

(iv) *The identity element* 1 *in A is homogeneous of degree* 0*.*

Proof. The proofs of (i) and (ii) are left as (routine) exercises.

(iii) If I is graded, then $I = \sum_p I^p$, so that I is generated by $\bigcup_p I^p$. But $\bigcup_p I^p$ consists of homogeneous elements because $I^p = I \cap A^p \subseteq A^p$ for all p.

Conversely, suppose that I is generated by a set X of homogeneous elements. We must show that $I = \sum_p (I \cap A_p)$, and it is only necessary to prove $I \subseteq \sum_p (I \cap A_p)$, for the reverse inclusion always holds. Since I is the two-sided ideal generated by X, a typical element $u \in I$ has the form $u = \sum_i a_i x_i b_i$, where $a_i, b_i \in A$ and $x_i \in X$. Now $u = \sum_p u_p$, where $u_p \in A^p$, and it suffices to show that each u_p lies in I. Indeed, it suffices to prove this for a single term $a_i x_i b_i$, and so we drop the subscript i. Since $a = \sum a_j$ and $b = \sum b_\ell$, where each a_j and b_ℓ are homogeneous, we have $u = \sum_{j,\ell} a_j x b_\ell$; but each

[16] There is a more general definition of a graded map $f : A \to B$. Given $d \in \mathbb{Z}$, then a k-algebra map f is ***graded map of degree d*** if $f(A^p) \subseteq B^{p+d}$ for all $p \geq 0$.

term in this sum is homogeneous, being the product of the homogeneous elements a_j, x, and b_ℓ. Thus, u_p is the sum of those $a_j x b_\ell$ having degree p, and so $u_p \in I$.

(iv) Write $1 = e_0 + e_1 + \cdots + e_t$, where $e_i \in A^i$. If $a_p \in A^p$, then

$$a_p - e_0 a_p = e_1 a_p + \cdots + e_t a_p \in A^p \cap (A^{p+1} \oplus \cdots \oplus A^{p+t}) = \{0\}.$$

It follows that $a_p = e_0 a_p$ for all homogeneous elements a_p, and so $a = e_0 a$ for all $a \in A$. A similar argument, examining $a_p = a_p 1$ (instead of $a_p = 1 a_p$), shows that $a = a e_0$ for all $a \in A$. Therefore, $1 = e_0$, by the uniqueness of the identity element in a ring. •

Example 9.96.
The quotient $R[x]/(x^{13})$ is a graded R-algebra. However, there is no obvious grading on the algebra $R[x]/(x^{13} + 1)$. After all, what degree should be assigned to the coset of x^{13}, which is the same as the coset of -1? ◂

We now consider generalized associativity of tensor product.

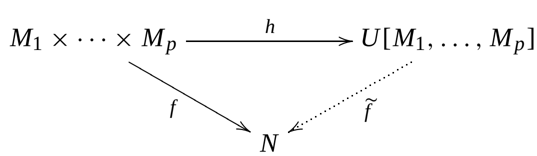
Definition. Let R be a commutative ring and let $M_1, \ldots, M_p$ be R-modules. An ***R-multilinear function*** $f: M_1 \times \cdots \times M_p \to N$, where N is an R-module, is a function that is additive in each of the p variables (when we fix the other $p - 1$ variables) and if $1 \le i \le p$, then

$$f(m_1, \ldots, r m_i, \ldots, m_p) = r f(m_1, \ldots, m_i, \ldots, m_p),$$

where $r \in R$ and $m_\ell \in M_\ell$ for all ℓ.

Proposition 9.97. *Let R be a commutative ring and let $M_1, \ldots, M_p$ be R-modules.*

(i) *There exists an R-module $U[M_1, \ldots, M_p]$ that is a solution to the universal mapping problem posed by multilinearity*:



There is a R-multilinear h such that, if f is R-multilinear, then there exists a unique R-homomorphism $\widetilde{f}$ making the diagram commute.

(ii) *If $f_i: M_i \to M_i'$ are R-maps, then there is a unique R-map*

$$u[f_1, \cdots, f_p]: U[M_1, \ldots, M_p] \to U[M_1', \ldots, M_p']$$

taking $h(m_1, \ldots, m_p) \mapsto h'(f_1(m_1), \ldots, f_p(m_p))$, where $h': M_1' \times \cdots \times M_p' \to U[M_1', \ldots, M_p']$.

Proof. (i) This is a straightforward generalization of Theorem 8.74, the existence of tensor products, using multilinear functions instead of bilinear ones. Let F be the free R-module with basis $M_1 \times \cdots \times M_p$, and let S be the submodule of F generated by all elements of the following two types:

$$(m_1, \ldots, m_i + m_i', \ldots, m_p) - (m_1, \ldots, m_i, \ldots, m_p) - (m_1, \ldots, m_i', \ldots, m_p);$$

$$(m_1, \ldots, rm_i, \ldots, m_p) - r(m_1, \ldots, m_i, \ldots, m_p),$$

where $m_i, m_i' \in M_i$, $r \in R$, and $1 \leq i \leq p$.

Define $U[M_1, \cdots, M_p] = F/S$ and define $h\colon M_1 \times \cdots \times M_p \to U[M_1, \cdots, M_p]$ by

$$h\colon (m_1, \ldots, m_p) \mapsto (m_1, \ldots, m_p) + S.$$

The reader should check that h is R-multilinear. The remainder of the proof is merely an adaptation of the proof of Proposition 8.74, and it is also left to the reader.

(ii) The function $M_1 \times \cdots \times M_p \to U[M_1', \ldots, M_p']$, given by

$$(m_1, \ldots, m_p) \mapsto h'(f_1(m_1), \ldots, f_n(m_p)),$$

is easily seen to be R-multilinear, and hence there exists a unique R-homomorphism as described in the statement. •

Observe that there are no parentheses needed in the generator $h(m_1, \ldots, m_p)$; that is, $h(m_1, \ldots, m_p)$ depends only on the p-tuple $(m_1, \ldots, m_p)$ and not on any association of its coordinates. The next proposition relates this construction to iterated tensor products. Once this is done, we will change the notation $U[M_1, \ldots, M_p]$.

Proposition 9.98 (Generalized Associativity). *Let R be a commutative ring and let $M_1, \ldots, M_p$ be R-modules. If $M_1 \otimes_R \cdots \otimes_R M_p$ is an iterated tensor product in some association, then there is an R-isomorphism $U[M_1, \ldots, M_p] \to M_1 \otimes_R \cdots \otimes_R M_p$ taking $h(m_1, \ldots, m_p) \mapsto m_1 \otimes \cdots \otimes m_p$.*

Remark. We are tempted to quote Theorem 2.20: Associativity for three factors implies associativity for many factors, for we have proved the associative law for three factors in Proposition 8.84. However, we did not prove equality, $A \otimes_R (B \otimes_R C) = (A \otimes_R B) \otimes_R C$; we only constructed an isomorphism. There is an extra condition, due, independently, to Mac Lane and Stasheff: If the associative law holds up to isomorphism and if a certain "pentagonal" diagram commutes, then generalized associativity holds up to isomorphism (see Mac Lane, *Categories for the Working Mathematician*, pages 157–161). ◂

Proof. The proof is by induction on $p \geq 2$. The base step is true, for $U[M_1, M_2] = M_1 \otimes_R M_2$. For the inductive step, let us assume that

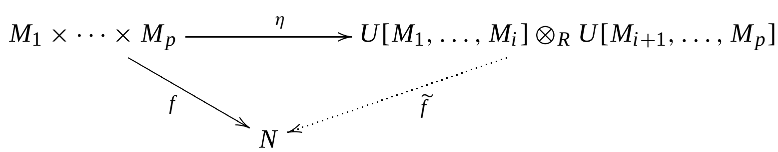
$$M_1 \otimes_R \cdots \otimes_R M_p = U[M_1, \ldots, M_i] \otimes_R U[M_{i+1}, \ldots, M_p].$$

We have indicated the final factors in the association; for example,

$$((M_1 \otimes_R M_2) \otimes_R M_3) \otimes_R (M_4 \otimes_R M_5) = U[M_1, M_2, M_3] \otimes_R U[M_4, M_5].$$

By induction, there are multilinear functions $h' \colon M_1 \times \cdots \times M_i \to M_1 \otimes_R \cdots \otimes_R M_i$ and $h'' \colon M_{i+1} \times \cdots \times M_p \to M_{i+1} \otimes_R \cdots \otimes_R M_p$ with $h'(m_1, \ldots, m_i) = m_1 \otimes \cdots \otimes m_i$ associated as in $M_1 \otimes_R \cdots \otimes_R M_i$, and with $h''(m_{i+1}, \ldots, m_p) = m_{i+1} \otimes \cdots \otimes m_p$ associated as in $M_{i+1} \otimes_R \cdots \otimes_R M_p$. Induction gives isomorphisms $\varphi' \colon U[M_1, \ldots, M_i] \to M_1 \otimes_R \cdots \otimes_R M_i$ and $\varphi'' \colon U[M_{i+1}, \ldots, M_p] \to M_{i+1} \otimes_R \cdots \otimes_R M_p$ with $\varphi' h' = h|(M_1 \times \cdots \times M_i)$ and $\varphi'' h'' = h|(M_{i+1} \times \cdots \times M_p)$. By Corollary 8.78, $\varphi' \otimes \varphi''$ is an isomorphism $U[M_1, \ldots, M_i] \otimes_R U[M_{i+1}, \ldots, M_p] \to M_1 \otimes_R \cdots \otimes_R M_p$.

We now show that $U[M_1, \ldots, M_i] \otimes_R U[M_{i+1}, \ldots, M_p]$ is a solution to the universal problem for multilinear functions. Consider the diagram



where $\eta(m_1, \ldots, m_p) = h'(m_1, \ldots, m_i) \otimes h''(m_{i+1}, \ldots, m_p)$, N is an R-module, and f is multilinear. We must find a homomorphism $\widetilde{f}$ making the diagram commute.

If $(m_1, \ldots, m_i) \in M_1 \times \cdots \times M_i$, the function $f_{(m_1,\ldots,m_i)} \colon M_{i+1} \times \cdots \times M_p \to N$, defined by $(m_{i+1}, \ldots, m_p) \mapsto f(m_1, \ldots, m_i, h''(m_{i+1}, \ldots, m_p))$, is multilinear; hence, there is a unique homomorphism $\widetilde{f}_{(m_1,\ldots,m_i)} \colon U[M_{i+1}, \ldots, M_p] \to N$ with

$$\widetilde{f}_{(m_1,\ldots,m_i)} \colon h''(m_{i+1}, \ldots, m_p) \mapsto f(m_1, \ldots, m_p).$$

If $r \in R$ and $1 \le j \le i$, then

$$\begin{aligned}
\widetilde{f}_{(m_1,\ldots,rm_j,\ldots,m_i)}(h''(m_{i+1}, \ldots, m_p)) &= f(m_1, \ldots, rm_j, \ldots, m_p) \\
&= rf(m_1, \ldots, m_j, \ldots, m_i) \\
&= r\,\widetilde{f}_{(m_1,\ldots,m_i)}(h''(m_{i+1}, \ldots, m_p)).
\end{aligned}$$

Similarly, if $m_j, m_j' \in M_j$, where $1 \le j \le i$, then

$$\widetilde{f}_{(m_1,\ldots,m_j+m_j',\ldots,m_i)} = \widetilde{f}_{(m_1,\ldots,m_j,\ldots,m_i)} + \widetilde{f}_{(m_1,\ldots,m_j',\ldots,m_i)}.$$

The function of $i+1$ variables $M_1 \times \cdots \times M_i \times U[M_{i+1}, \ldots, M_p] \to N$, defined by $(m_1, \ldots, m_i, u'') \mapsto \widetilde{f}_{(m_1,\ldots,m_i)}(u'')$, is multilinear, and so it gives a bilinear function $U[M_1, \ldots, M_i] \times U[M_{i+1}, \ldots, M_p] \to N$, namely, $(u', u'') \mapsto (h'(u'), h''(u''))$. Thus, there is a unique homomorphism $\widetilde{f} \colon U[M_1, \ldots, M_i] \otimes_R U[M_{i+1}, \ldots, M_p] \to N$ which takes $h'(m_1, \ldots, m_i) \otimes h''(m_{i+1}, \ldots, m_p) \mapsto \widetilde{f}_{(m_1,\ldots,m_i)}(h''(m_{i+1}, \ldots, m_p)) = f(m_1, \ldots, m_p)$; that is, $\widetilde{f}\eta = f$. Therefore, $U[M_1, \ldots, M_i] \otimes_R U[M_{i+1}, \ldots, M_p]$ is a solution to the universal mapping problem. By uniqueness of such solutions, there is an isomorphism $\theta \colon U[M_1, \ldots, M_p] \to U[M_1, \ldots, M_i] \otimes_R U[M_{i+1}, \ldots, M_p]$ with $\theta h(m_1, \ldots, m_p) = h'(m_1, \ldots, m_i) \otimes h''(m_{i+1}, \ldots, m_p) = \eta(m_1, \ldots, m_p)$. Finally, $(\varphi' \otimes \varphi'')\theta$ is the desired isomorphism $U[M_1, \ldots, M_p] \cong M_1 \otimes_R \cdots \otimes_R M_p$. •

We now abandon the notation in Proposition 9.97; from now on, we shall write

$$U[M_1, \dots, M_p] = M_1 \otimes_R \cdots \otimes_R M_p,$$
$$h(m_1, \dots, m_p) = m_1 \otimes \cdots \otimes m_p,$$
$$u[f_1, \dots, f_p] = f_1 \otimes \cdots \otimes f_p.$$

Proposition 9.99. *If R is a commutative ring and A and B are R-algebras, then the tensor product $A \otimes_R B$ is an R-algebra if we define $(a \otimes b)(a' \otimes b') = aa' \otimes bb'$.*

Proof. First, $A \otimes_R B$ is an R-module, by Corollary 8.81. Let $\mu \colon A \times A \to A$ and $\nu \colon B \times B \to B$ be the given multiplications on the algebras A and B, respectively. We must show there is a multiplication on $A \otimes_R B$ as in the statement; that is, there is an R-bilinear function $\lambda \colon (A \otimes_R B) \times (A \otimes_R B) \to A \otimes_R B$ with $\lambda \colon (a \otimes b, a' \otimes b') \mapsto aa' \otimes bb'$. Such a function λ exists because it is the composite

$$\begin{aligned}(A \otimes_R B) \times (A \otimes_R B) &\to (A \otimes_R B) \otimes (A \otimes_R B)\\ &\to (A \otimes_R A) \times (B \otimes_R B)\\ &\to A \otimes_R B :\end{aligned}$$

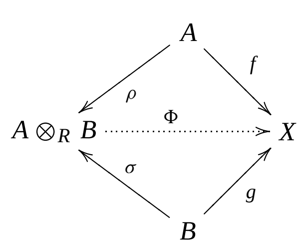
the first function is $(a \otimes b, a' \otimes b') \mapsto a \otimes b \otimes a' \otimes b'$ (which is the bilinear function in Proposition 8.82); the second is $1 \otimes \tau \otimes 1$, where $\tau \colon B \otimes_R A \to A \otimes_R B$ takes $b \otimes a \mapsto a \otimes b$ (which exists by Propositions 8.83 and 9.98); the third is $\mu \otimes \nu$. It is now routine to check that the R-module $A \otimes_R B$ is an R-algebra. •

Example 9.100.
In Exercise 8.48 on page 604, we saw that there is an isomorphism of abelian groups: $\mathbb{I}_m \otimes \mathbb{I}_n \cong \mathbb{I}_d$, where $d = (m, n)$. It follows that if $(m, n) = 1$, then $\mathbb{I}_m \otimes \mathbb{I}_n = \{0\}$. Of course, this tensor product is still $\{0\}$ if we regard $\mathbb{I}_m$ and $\mathbb{I}_n$ as $\mathbb{Z}$-algebras. Thus, in this case, the tensor product is the zero ring. Had we insisted, in the definition of ring, that $1 \neq 0$, then the tensor product of rings would not always be defined. ◂

We now show that the tensor product of algebras is an "honest" construction.

Proposition 9.101. *If R is a commutative ring and A and B are commutative R-algebras, then $A \otimes_R B$ is the coproduct in the category of commutative R-algebras.*

Proof. Define $\rho \colon A \to A \otimes_R B$ by $\rho \colon a \mapsto a \otimes 1$, and define $\sigma \colon B \to A \otimes_R B$ by $\sigma : b \mapsto 1 \otimes b$. Let X be a commutative R-algebra, and consider the diagram



where f and g are R-algebra maps. The function $\varphi\colon A \times B \to X$, given by $(a, b) \mapsto f(a)g(b)$, is easily seen to be R-bilinear, and so there is a unique map of R-modules $\Phi\colon A \otimes_R B \to X$ with $\Phi(a \otimes b) = f(a)g(b)$. It remains to prove that Φ is an R-algebra map, for which it suffices to prove that $\Phi\big((a \otimes b)(a' \otimes b')\big) = \Phi(a \otimes b)\Phi(a' \otimes b')$. Now

$$\begin{aligned}\Phi\big((a \otimes b)(a' \otimes b')\big) &= \Phi(aa' \otimes bb')\\ &= f(a)f(a')g(b)g(b').\end{aligned}$$

On the other hand, $\Phi(a \otimes b)\Phi(a' \otimes b') = f(a)g(b)f(a')g(b')$. Since X is commutative, Φ does preserve multiplication. •

Bimodules can be viewed as left modules over a suitable ring.

Corollary 9.102. *Let R and S be k-algebras, where k is a commutative ring. Every (R, S)-bimodule M is a left $R \otimes_k S^{\mathrm{op}}$-module, where*

$$(r \otimes s)m = rms.$$

Proof. The function $R \times S^{\mathrm{op}} \times M \to M$, given by $(r, s, m) \mapsto rms$, is k-trilinear, and this can be used to prove that $(r \otimes s)m = rms$ is well-defined. Let us write $s * s'$ for the product in S^{op}; that is, $s * s' = s's$. The only axiom that is not obvious is axiom (iii) in the definition of module: If $a, a' \in R \otimes_k S^{\mathrm{op}}$, then $(aa')m = a(a'm)$, and it is enough to check that this is true for generators $a = r \otimes s$ and $a' = r' \otimes s'$ of $R \otimes_k S^{\mathrm{op}}$. But

$$\begin{aligned}[(r \otimes s)(r' \otimes s')]m &= [rr' \otimes s * s']m\\ &= (rr')m(s * s')\\ &= (rr')m(s's)\\ &= r(r'ms')s.\end{aligned}$$

On the other hand,

$$(r \otimes s)[(r' \otimes s')m] = (r \otimes s)[r'(ms')] = r(r'ms')s. \quad \bullet$$

Definition. If k is a commutative ring and A is a k-algebra, then its ***enveloping algebra*** is

$$A^e = A \otimes_k A^{\mathrm{op}}.$$

Corollary 9.103. *If k is a commutative ring and A is a k-algebra, then A is a left A^e-module whose submodules are the two-sided ideals. If A is a simple k-algebra, then A is a simple A^e-module.*

Proof. Since a k-algebra A is an (A, A)-bimodule, it is a left A^e-module. •

Proposition 9.104. *If k is a commutative ring and A is a k-algebra, then*

$$\mathrm{End}_{A^e}(A) \cong Z(A).$$

Proof. If $f\colon A \to A$ is an A^e-map, then it is a map of A viewed only as a left A-module. Proposition 8.12 applies to say that f is determined by $z = f(1)$, because $f(a) = f(a1) = af(1) = az$ for all $a \in A$. On the other hand, since f is also a map of A viewed as a right A-module, we have $f(a) = f(1a) = f(1)a = za$. Therefore, $z = f(1) \in Z(A)$; that is, the map $\varphi\colon f \mapsto f(1)$ is a map $\mathrm{End}_{A^e}(A) \to Z(A)$. The map φ is surjective, for if $z \in Z(A)$, then $f(a) = za$ is an A^e-endomorphism with $\varphi(f) = z$; the map φ is injective, for if $f \in \mathrm{End}_{A^e}(A)$ and $f(1) = 0$, then $f = 0$. •

We now construct the *tensor algebra* on an R-module M. When M is a free R-module with basis X, then the tensor algebra will be seen to be the free R-algebra with basis X; that is, it is the polynomial ring over R in noncommuting variables X.

Definition. Let R be a commutative ring, and let M be an R-module. Define

$$\begin{aligned} &T^0(M) = R,\\ &T^1(M) = M,\\ &T^p(M) = M \otimes_R \cdots \otimes_R M \quad (p \text{ times}) \quad \text{if } p \geq 2. \end{aligned}$$

Remark. Many authors denote $T^p(M)$ by $\bigotimes^p M$. In Proposition 9.97, $T^p(M)$ was originally denoted by $U[M_1, \ldots, M_p]$ (here, all $M_i = M$), and we later replaced this notation by $M_1 \otimes \cdots \otimes M_p$, for this is easier to remember. We remind the reader that $T^p(M)$, however it is denoted, is generated by symbols $m_1 \otimes \cdots \otimes m_p$ in which no parentheses occur. ◂

Proposition 9.105. *If M is an R-module, then there is a graded R-algebra*

$$T(M) = \sum_{p \geq 0} T^p(M)$$

with the action of $r \in R$ on $T^q(M)$ given by

$$r(y_1 \otimes \cdots \otimes y_q) = (ry_1) \otimes y_2 \otimes \cdots \otimes y_q = (y_1 \otimes \cdots \otimes y_q)r,$$

and with the multiplication $T^p(M) \times T^q(M) \to T^{p+q}(M)$, for $p, q \geq 1$, given by

$$(x_1 \otimes \cdots \otimes x_p, y_1 \otimes \cdots \otimes y_q) \mapsto x_1 \otimes \cdots \otimes x_p \otimes y_1 \otimes \cdots \otimes y_q.$$

Proof. First, define the product of two homogeneous elements by the formulas in the statement. Multiplication $\mu\colon T(M) \times T(M) \to T(M)$ must now be

$$\mu\colon \Big(\sum_p x_p, \sum_q y_q\Big) \mapsto \sum_{p,q} x_p \otimes y_q,$$

where $x_p \in T^p(M)$ and $y_q \in T^q(M)$. Multiplication is associative because no parentheses are needed in describing generators $m_1 \otimes \cdots \otimes m_p$ of $T^p(M)$, and the distributive laws hold because multiplication is R-bilinear. Finally, $1 \in k = T^0(M)$ is the identity, each element of R commutes with every element of $T(M)$, and $T^p(M)T^q(M) \subseteq T^{p+q}(M)$, so that $T(M)$ is a graded R-algebra. •

The reader may check that if $M = R$, then $T(M) \cong R[x]$.

Definition. If R is a commutative ring and M is an R-module, then $T(M)$ is called the ***tensor algebra*** on M.

If R is a commutative ring and A and B are R-modules, define a ***word*** of ***length*** $p \geq 0$ on A and B to be an R-module of the form

$$W(A, B)_p = T^{e_1}(A) \otimes_R T^{f_1}(B) \otimes_R \cdots \otimes_R T^{e_r}(A) \otimes_R T^{f_r}(B),$$

where $\sum_i (e_i + f_i) = p$, all $e_i,\ f_i$ are integers, $e_1 \geq 0,\ f_r \geq 0$, and all the other exponents are positive.

Proposition 9.106. *If A and B are R-modules, then for all $p \geq 0$,*

$$T^p(A \oplus B) \cong \sum_{j=0}^{p} W(A, B)_j \otimes_R W'(A, B)_{p-j},$$

where $W(A, B)_j$, $W'(A, B)_{p-j}$ range over all words of length j and $p - j$, respectively.

Proof. The proof is by induction on $p \geq 0$. For the base step,

$$T^0(A \oplus B) = R \cong R \otimes_R R \cong T^0(A) \otimes_R T^0(B).$$

For the inductive step,

$$\begin{aligned} T^{p+1}(A \oplus B) &= T^p(A \oplus B) \otimes_R (A \oplus B) \\ &\cong \big(T^p(A \oplus B) \otimes_R A\big) \oplus \big(T^p(A \oplus B) \otimes_R B\big) \\ &\cong \sum_{j=0}^{p} W(A, B)_j \otimes_R W'(A, B)_{p-j} \otimes_R X, \end{aligned}$$

where $X \cong A$ or $X \cong B$. This completes the proof, for every word of length $p - j + 1$ has the form $W'(A, B) \otimes_R X$. •

Proposition 9.107. *Tensor algebra defines a functor $T\colon {}_R\mathbf{Mod} \to \mathbf{Gr}_R\mathbf{Alg}$. Moreover, T preserves surjections.*

Proof. We have already defined T on every R-module M: it is the tensor algebra $T(M)$. If $f\colon M \to N$ is an R-homomorphism, then Proposition 9.97 provides maps

$$f \otimes \cdots \otimes f\colon T^p(M) \to T^p(N),$$

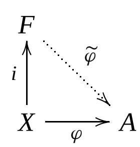
for each p, which give an R-algebra map $T(M) \to T(N)$. It is a simple matter to check that T preserves identity maps and composites.

Assume that $f\colon M \to N$ is a surjective R-map. If $n_1 \otimes \cdots \otimes n_p \in T^p(N)$, then surjectivity of f provides $m_i \in M$, for all i, with $f(m_i) = n_i$, and so

$$T(f)\colon m_1 \otimes \cdots \otimes m_p \mapsto n_1 \otimes \cdots \otimes n_p. \quad \bullet$$

We now generalize the notion of free module to free algebra.

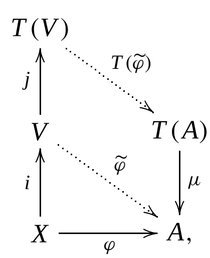
Definition. If X is a subset of an R-algebra F, then F is a ***free R-algebra*** with ***basis*** X if, for every R-algebra A and every function $\varphi\colon X \to A$, there exists a unique R-algebra map $\widetilde{\varphi}$ with $\widetilde{\varphi}(x) = \varphi(x)$ for all $x \in X$. In other words, the following diagram commutes, where $i\colon X \to F$ is the inclusion.



In the next proposition, we regard the graded R-algebra $T(V)$ merely as an R-algebra.

Proposition 9.108. *If V is a free R-module with basis X, where R is a commutative ring, then $T(V)$ is a free R-algebra with basis X.*

Proof. Consider the diagram



where $i\colon X \to V$ and $j\colon V \to T(V)$ are inclusions, and A is an R-algebra. Viewing A only as an R-module gives an R-module map $\widetilde{\varphi}\colon V \to A$, for V is a free R-module

with basis X. Applying the functor T gives an R-algebra map $T(\widetilde{\varphi})\colon T(V) \to T(A)$. For existence of an R-algebra map $T(V) \to A$, it suffices to define an R-algebra map $\mu\colon T(A) \to A$ such that the composite $\mu \circ T(\widetilde{\varphi})$ is an R-algebra map extending φ. For each p, consider the diagram

$$\begin{array}{ccc} A \times \cdots \times A & \xrightarrow{h_p} & T^p(A) \\ & \searrow^{m_p} & \downarrow{\scriptstyle \mu_p} \\ & & A, \end{array}$$

where $h_p\colon (a_1, \ldots, a_p) \mapsto a_1 \otimes \cdots \otimes a_p$ and $m_p\colon (a_1, \ldots, a_p) \mapsto a_1 \cdots a_p$, the latter being the product of the elements $a_1, \ldots, a_p$ in the R-algebra A. Of course, m_p is R-multilinear, and so it induces an R-map μ_p making the diagram commute. Now define $\mu\colon T(A) \to A$ by $\mu = \sum_p \mu_p$. To see that μ is multiplicative, it suffices to show

$$\mu_{p+q}\big((a_1 \otimes \cdots \otimes a_p) \otimes (a_1' \otimes \cdots \otimes a_q')\big) = \mu_p(a_1 \otimes \cdots \otimes a_p)\mu_q(a_1' \otimes \cdots \otimes a_q').$$

But this equation follows from the associative law in A:

$$(a_1 \cdots a_p)(a_1' \cdots a_q') = a_1 \cdots a_p a_1' \cdots a_q'.$$

Finally, uniqueness of this R-algebra map follows from V generating $T(V)$ as an R-algebra [after all, every homogeneous element in $T(V)$ is a product of elements of degree 1]. •

Corollary 9.109. *Let R be a commutative ring.*

(i) *If A is an R-algebra, then there is a surjective R-algebra map $T(A) \to A$.*

(ii) *Every R-algebra A is a quotient of a free R-algebra.*

Proof. (i) Regard A only as an R-module. For each $p \geq 2$, multiplication $A^p \to A$ is R-multilinear, and so there is a unique R-module map $T^p(A) \to A$. But these maps may be assembled to give an R-module map $T(A) = \sum_p T^p(A) \to A$. This map is surjective, because A has a unit 1, and it is easily seen to be a map of R-algebras; that is, it preserves multiplication.

(ii) Let V be a free R-module for which there exists a surjective R-map $\varphi\colon V \to A$. By Proposition 9.107, the induced map $T(\varphi)\colon T(V) \to T(A)$ is surjective. Now $T(V)$ is a free R-algebra, and if we compose $T(\varphi)$ with the surjection $T(A) \to A$, then A is a quotient of $T(V)$. •

Definition. If R is a commutative ring and V is a free R-module with basis X, then $T(V)$ is called the ring of polynomials over R in ***noncommuting variables*** X, and it is denoted by $R\langle X\rangle$.

If V is the free R-module with basis X, then each element u in $T(V)$ has a unique expression

$$u = \sum_{\substack{p \geq 0 \\ i_1, \ldots, i_p}} r_{i_1, \ldots, i_p} x_{i_1} \otimes \cdots \otimes x_{i_p},$$

where $r_{i_1, \ldots, i_p} \in R$ and $x_{i_j} \in X$. We obtain the usual notation for such a polynomial by erasing the tensor product symbols. For example, if $X = \{x, y\}$, then

$$u = r_0 + r_1 x + r_2 y + r_3 x^2 + r_4 y^2 + r_5 xy + r_6 yx + \cdots .$$

Example 9.110.
Just as for modules, we can now construct rings ($\mathbb{Z}$-algebras) by generators and relations. The first example of a ring that is left noetherian but not right noetherian was given by J. Dieudonné; it is the ring R generated by elements x and y satisfying the relations $yx = 0$ and $y^2 = 0$. The existence of the ring R is now easy: Let V be the free abelian group with basis u, v, let $R = \left(\sum_{p \geq 0} T^p(V)\right)/I$, where I is the two-sided ideal generated by vu and v^2, and set $x = u + I$ and $y = v + I$. Note that since the ideal I is generated by homogeneous elements of degree 2, we have $T^1(V) = V \cap I = \{0\}$, and so $x \neq 0$ and $y \neq 0$. ◂

We now mention a class of rings generalizing commutative rings.

Definition. If k is a field,[17] then a ***polynomial identity*** on a k-algebra A is an element $f(X) \in k\langle X \rangle$ (the ring of polynomials over k in noncommuting variables X) all of whose substitutions in A are 0.

For example, if $f(x, y) = xy - yx \in k\langle x, y \rangle$, then f is a polynomial identity on a k-algebra A if $ab - ba = 0$ for all $a, b \in A$; that is, A is commutative.

Here is a precise definition. Every function $\varphi \colon X \to A$ extends to a k-algebra map $\widetilde{\varphi} \colon k\langle X \rangle \to A$, and $f(X)$ is a polynomial identity on A if and only if $f(X) \in \bigcap_{\varphi} \ker \widetilde{\varphi}$ for all functions $\varphi \colon X \to A$.

Definition. A k-algebra A is a ***PI-algebra*** (an algebra satisfying a polynomial identity) if A satisfies some identity at least one of whose coefficients is 1.

Every k-algebra generated by n elements satisfies the ***standard identity***

$$s_{n+1}(x_1, \ldots, x_{n+1}) = \sum_{\sigma \in S_{n+1}} \operatorname{sgn}(\sigma) x_{\sigma(1)} \cdots x_{\sigma(n+1)}.$$

We can prove that the matrix algebra $\mathrm{Mat}_n(k)$ satisfies the standard identity s_{n^2+1}, and S. A. Amitsur and J. Levitzki proved that $\mathrm{Mat}_n(k)$ satisfies s_{2n}; moreover, $2n$ is the lowest possible degree of such a polynomial identity. There is a short proof of this due to S. Rosset, "A New Proof of the Amitsur-Levitski Identity," *Israel Journal of Mathematics* 23, 1976, pages 187–188.

[17] We could, of course, extend these definitions by allowing k to be a commutative ring.

Definition. A ***central polynomial identity*** on a k-algebra A is a polynomial identity $f(X) \in k\langle X\rangle$ on A all of whose values $f(a_1, a_2, \ldots)$ (as the a_i vary over all elements of A) lie in $Z(A)$.

It was proved, independently, by E. Formanek and Yu. P. Razmyslov, that $\mathrm{Mat}_n(k)$ satisfies a central polynomial identity.

There are theorems showing, in several respects, that PI-algebras behave like commutative algebras. For example, recall that a ring R is *primitive* if it has a faithful simple left R-module; if R is commutative, then R is a field. I. Kaplansky proved that every primitive quotient of a PI-algebra is simple and finite-dimensional over its center. The reader is referred to Procesi, *Rings with Polynomial Identities*.

Another interesting area of current research involves *noncommutative algebraic geometry*. In essence, this involves the study of varieties now defined as zeros of ideals in $k\langle x_1, \ldots, x_n\rangle$ instead of in $k[x_1, \ldots, x_n]$.

EXERCISES

9.62 (i) If k is a subfield of a field K, prove that the ring $K \otimes_k k[x]$ is isomorphic to $K[x]$.

(ii) Suppose that k is a field, $p(x) \in k[x]$ is irreducible, and $K = k(\alpha)$, where α is a root of $p(x)$. Prove that, as rings, $K \otimes_k K \cong K[x]/(p(x))$, where $(p(x))$ is the principal ideal in $K[x]$ generated by $p(x)$.

(iii) The polynomial $p(x)$, though irreducible in $k[x]$, may factor in $K[x]$. Give an example showing that the ring $K \otimes_k K$ need not be semisimple.

(iv) Prove that if K/k is a finite separable extension, then $K \otimes_k K$ is semisimple. (The converse is also true.)

9.63 Let m and n be positive integers, and let $d = \gcd(m, n)$. Prove that $\mathbb{I}_m \otimes_{\mathbb{Z}} \mathbb{I}_n \cong \mathbb{I}_d$ as commutative rings.

Hint. See Exercise 8.48 on page 604.

9.64 If $A \cong A'$ and $B \cong B'$ are k-algebras, where k is a commutative ring, prove that $A \otimes_k B \cong A' \otimes_k B'$ as k-algebras.

9.65 If k is a commutative ring and A and B are k-algebras, prove that

$$(A \otimes_k B)^{\mathrm{op}} \cong A^{\mathrm{op}} \otimes_k B^{\mathrm{op}}.$$

9.66 If R is a commutative k-algebra, where k is a field, and if G is a group, prove that $R \otimes_k kG \cong RG$.

9.67 (i) If k is a subring of a commutative ring R, prove that $R \otimes_k k[x] \cong R[x]$ as R-algebras.

(ii) If $f(x) \in k[x]$ and (f) is the principal ideal in $k[x]$ generated by $f(x)$, prove that $R \otimes_k (f)$ is the principal ideal in $R[x]$ generated by $f(x)$. More precisely, there is a commutative diagram

$$\begin{array}{ccccc} 0 & \longrightarrow & E \otimes_k (f) & \longrightarrow & E \otimes_k k[x] \\ & & \downarrow & & \downarrow \\ 0 & \longrightarrow & (f)_E & \longrightarrow & E[x] \end{array}$$

(iii) Let k be a field and $E \cong k[x]/(f)$, where $f(x) \in k[x]$ is irreducible. Prove that $E \otimes_k E \cong E[x]/(f)_E$, where $(f)_E$ is the principal ideal in $E[x]$ generated by $f(x)$.

(iv) Give an example of a field extension E/k with $E \otimes_k E$ not a field.

Hint. If $f(x) \in k[x]$ factors into $g(x)h(x)$ in $E[x]$, where $(g, h) = 1$, then the Chinese remainder theorem applies.

9.68 Let k be a field and let $f(x) \in k[x]$ be irreducible. If K/k is a field extension, then $f(x) = p_1(x)^{e_1} \cdots p_n(x)^{e_n} \in K[x]$, where the $p_i(x)$ are distinct irreducible polynomials in $K[x]$ and $e_i \geq 1$.

(i) Prove that $f(x)$ is separable if and only if all $e_i = 1$.

(ii) Prove that a finite field extension K/k is separable if and only if $K \otimes_k K$ is a semisimple ring.

Hint. First, observe that K/k is a simple extension, so there is an exact sequence $0 \to (f) \to k[x] \to K \to 0$. Second, use the Chinese remainder theorem.

9.69 Prove that the ring R in Example 9.110 is left noetherian but not right noetherian.
Hint. See Cartan and Eilenberg, *Homological Algebra*, p. 16.

9.70 If G is a group, then a k-algebra A is called ***G-graded*** if there are k-submodules A^g, for all $g \in G$, such that

(i) $A = \sum_{g \in G} A^g$;

(ii) For all $g, h \in G$, $A^g A^h \subseteq A^{gh}$.

An $\mathbb{I}_2$-graded algebra is called a ***superalgebra***. If A is a G-graded algebra and e is the identity element of G, prove that $1 \in A^e$.

9.71 If A is a k-algebra generated by n elements, prove that A satisfies the standard identity defined on page 725.

9.7 DIVISION ALGEBRAS

That the tensor product of algebras is, again, an algebra, is used in the study of division rings.

Definition. A ***division algebra*** over a field k is a division ring regarded as an algebra over its center k.

Let us begin by considering the wider class of simple algebras.

Definition. A k-algebra A over a field k is ***central simple*** if it is finite-dimensional,[18] simple (no two-sided ideals other than A and $\{0\}$), and its center $Z(A) = k$.

Notation. If A is an algebra over a field k, then we write

$$[A : k] = \dim_k(A).$$

[18] Some authors do not assume finite-dimensionality.

Example 9.111.

(i) Every division algebra Δ that is finite-dimensional over its center k is a central simple k-algebra. The quaternions $\mathbb{H}$ is a central simple $\mathbb{R}$-algebra, and every field is a central simple algebra over itself. Hilbert gave an example of an infinite-dimensional division algebra (see Drozd–Kirichenko, *Finite Dimensional Algebras*, page 81).

(ii) If k is a field, then $\mathrm{Mat}_n(k)$ is a central simple k-algebra.

(iii) If A is a central simple k-algebra, then its opposite algebra A^{op} is also a central simple k-algebra. ◂

Theorem 9.112. *Let A be a central simple k-algebra. If B is a simple k-algebra, then $A \otimes_k B$ is a central simple $Z(B)$-algebra. In particular, if B is a central simple k-algebra, then $A \otimes_k B$ is a central simple k-algebra.*

Proof. Each $x \in A \otimes_k B$ has an expression of the form

$$x = a_1 \otimes b_1 + \cdots + a_n \otimes b_n, \tag{1}$$

where $a_i \in A$ and $b_i \in B$. For nonzero x, define the *length* of x to be n if there is no such expression having fewer than n terms. We claim that if x has length n, that is, if Eq. (1) is a shortest such expression, then $b_1, \ldots, b_n$ is a linearly independent list in B (viewed as a vector space over k). Otherwise, there is some j and $u_i \in k$, not all zero, with $b_j = \sum_i u_i b_i$. Substituting and collecting terms gives

$$x = \sum_{i \neq j} (a_i + u_i a_j) \otimes b_i,$$

which is a shorter expression for x.

Let $I \neq \{0\}$ be a two-sided ideal in $A \otimes_k B$. Choose x to be a (nonzero) element in I of smallest length, and assume that Eq. (1) is a shortest expression for x. Now $a_1 \neq 0$. Since Aa_1A is a two-sided ideal in A, simplicity gives $A = Aa_1A$. Hence, there are elements a'_p and a''_p in A with $1 = \sum_p a'_p a_1 a''_p$. Since I is a two-sided ideal,

$$x' = \sum_p a'_p x a''_p = 1 \otimes b_1 + c_2 \otimes b_2 + \cdots + c_n \otimes b_n \tag{2}$$

lies in I, where, for $i \geq 2$, we have $c_i = \sum_p a'_p a_i a''_p$. At this stage, we do not know whether $x' \neq 0$, but we do know, for every $a \in A$, that $(a \otimes 1)x' - x'(a \otimes 1) \in I$. Now

$$(a \otimes 1)x' - x'(a \otimes 1) = \sum_{i \geq 2} (ac_i - c_i a) \otimes b_i. \tag{3}$$

First, this element is 0, lest it be an element in I of length smaller than the length of x. Since $b_1, \ldots b_n$ is a linearly independent list, the k-subspace it generates is $\langle b_1, \ldots, b_n \rangle = \langle b_1 \rangle \oplus \cdots \oplus \langle b_n \rangle$, and so

$$A \otimes_k \langle b_1, \ldots, b_n \rangle = A \otimes_k \langle b_1 \rangle \oplus \cdots \oplus A \otimes_k \langle b_n \rangle.$$

It follows from Eq. (3) that each term $(ac_i - c_i a) \otimes b_i$ must be 0. Hence, $ac_i = c_i a$ for all $a \in A$; that is, each $c_i \in Z(A) = k$. Eq. (2) becomes

$$\begin{aligned} x' &= 1 \otimes b_1 + c_2 \otimes b_2 + \cdots + c_n \otimes b_n \\ &= 1 \otimes b_1 + 1 \otimes c_2 b_2 + \cdots + 1 \otimes c_n b_n \\ &= 1 \otimes (b_1 + c_2 b_2 + \cdots + c_n b_n). \end{aligned}$$

Now $b_1 + c_2 b_2 + \cdots + c_n b_n \neq 0$, because $b_1, \ldots, b_n$ is a linearly independent list, and so $x' \neq 0$. Therefore, I contains a nonzero element of the form $1 \otimes b$. But simplicity of B gives $BbB = B$, and so there are $b'_q, b''_q \in B$ with $\sum_q b'_q b b''_q = 1$. Hence, I contains $\sum_q (1 \otimes b'_q)(1 \otimes b)(1 \otimes b''_q) = 1 \otimes 1$, which is the unit in $A \otimes_k B$. Therefore, $I = A \otimes_k B$ and $A \otimes_k B$ is simple.

We now seek the center of $A \otimes_k B$. Clearly, $k \otimes_k Z(B) \subseteq Z(A \otimes_k B)$. For the reverse inequality, let $z \in Z(A \otimes_k B)$ be nonzero, and let

$$z = a_1 \otimes b_1 + \cdots + a_n \otimes b_n$$

be a shortest such expression for z. As in the preceding argument, $b_1, \ldots, b_n$ is a linearly independent list over k. For each $a \in A$, we have

$$0 = (a \otimes 1)z - z(a \otimes 1) = \sum_i (aa_i - a_i a) \otimes b_i.$$

It follows, as above, that $(aa_i - a_i a) \otimes b_i = 0$ for each i. Hence, $aa_i - a_i a = 0$, so that $aa_i = a_i a$ for all $a \in A$ and each $a_i \in Z(A) = k$. Thus, $z = 1 \otimes x$, where $x = a_1 b_1 + \cdots + a_n b_n \in B$. But if $b \in B$, then

$$0 = z(1 \otimes b) - (1 \otimes b)z = (1 \otimes x)(1 \otimes b) - (1 \otimes b)(1 \otimes x) = 1 \otimes (xb - bx).$$

Therefore, $xb - bx = 0$ and $x \in Z(B)$. We conclude that $z \in k \otimes_k Z(B)$, as desired. •

It is not generally true that the tensor product of simple k-algebras is again simple; we must pay attention to the centers. In Exercise 9.67(iv) on page 727, we saw that if E/k is a field extension, then $E \otimes_k E$ need not be a field. The tensor product of division algebras need not be a division algebra, as we see in the next example.

Example 9.113.
The algebra $\mathbb{C} \otimes_{\mathbb{R}} \mathbb{H}$ is an eight-dimensional $\mathbb{R}$-algebra, but it is also a four-dimensional $\mathbb{C}$-algebra: A basis is

$$1 = 1 \otimes 1, \quad 1 \otimes i, \quad 1 \otimes j, \quad 1 \otimes k.$$

We let the reader prove that the vector space isomorphism $\mathbb{C} \otimes_{\mathbb{R}} \mathbb{H} \to \mathrm{Mat}_2(\mathbb{C})$ with

$$1 \otimes 1 \mapsto \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix},$$
$$1 \otimes i \mapsto \begin{bmatrix} i & 0 \\ 0 & -i \end{bmatrix},$$
$$1 \otimes j \mapsto \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix},$$
$$1 \otimes k \mapsto \begin{bmatrix} 0 & i \\ i & 0 \end{bmatrix},$$

is an isomomorphism of $\mathbb{C}$-algebras. ◀

Another way to see that $\mathbb{C} \otimes_{\mathbb{R}} \mathbb{H} \cong \mathrm{Mat}_2(\mathbb{C})$ arises from Example 8.71(ii). We remarked then that

$$\mathbb{R}\mathbf{Q} \cong \mathbb{R} \times \mathbb{R} \times \mathbb{R} \times \mathbb{R} \times \mathbb{H};$$

tensoring by $\mathbb{C}$ gives

$$\mathbb{C}\mathbf{Q} \cong \mathbb{C} \otimes_{\mathbb{R}} \mathbb{R}\mathbf{Q} \cong \mathbb{C} \times \mathbb{C} \times \mathbb{C} \times \mathbb{C} \times \mathbb{C} \otimes_{\mathbb{R}} \mathbb{H}.$$

It follows from the uniqueness in Wedderburn's theorem that $\mathbb{C} \otimes_{\mathbb{R}} \mathbb{H} \cong \mathrm{Mat}_2(\mathbb{C})$.

The next theorem puts the existence of the isomorphism in Example 9.113 into the context of central simple algebras.

Theorem 9.114. *Let k be a field and let A be a central simple k-algebra.*

(i) *If $\overline{k}$ is the algebraic closure of k, then there is an integer n with*

$$\overline{k} \otimes_k A \cong \mathrm{Mat}_n(\overline{k}).$$

(ii) *If A is a central simple k-algebra, then there is an integer n with*

$$[A : k] = n^2.$$

Proof. (i) By Theorem 9.112, $\overline{k} \otimes_k A$ is a simple $\overline{k}$-algebra. Hence, Wedderburn's theorem (actually, Corollary 8.63) gives $\overline{k} \otimes_k A \cong \mathrm{Mat}_n(D)$ for some $n \geq 1$ and some division ring D. Since D is a finite-dimensional division algebra over $\overline{k}$, the argument in Molien's Corollary 8.65 shows that $D = \overline{k}$.

(ii) We claim that $[A : k] = [\overline{k} \otimes_k A : \overline{k}]$, for if $a_1, \ldots, a_m$ is a basis of A over k, then $1 \otimes a_1, \ldots, 1 \otimes a_m$ is a basis of $\overline{k} \otimes_k A$ over $\overline{k}$ (essentially because tensor product commutes with direct sum). Therefore,

$$[A : k] = [\overline{k} \otimes_k A : \overline{k}] = [\mathrm{Mat}_n(\overline{k}) : \overline{k}] = n^2. \quad \bullet$$

The division ring of quaternions $\mathbb{H}$ is a central simple $\mathbb{R}$-algebra, and so its dimension $[\mathbb{H} : \mathbb{R}]$ must be a square (it is 4). Moreover, since $\mathbb{C}$ is algebraically closed, Theorem 9.114 gives $\mathbb{C} \otimes_{\mathbb{R}} \mathbb{H} \cong \mathrm{Mat}_2(\mathbb{C})$ (Example 9.113 displays an explicit isomorphism).

Definition. A ***splitting field*** for a central simple k-algebra A is a field extension E/k for which there exists an integer n such that $E \otimes_k A \cong \mathrm{Mat}_n(E)$.

Theorem 9.114 says that the algebraic closure $\overline{k}$ of a field k is a splitting field for every central simple k-algebra A. We are going to see that there always exists a splitting field that is a finite extension of k, but we first develop some tools in order to prove it.

Definition. If A is a k-algebra and $X \subseteq A$ is a subset, then its ***centralizer***, $C_A(X)$, is defined by

$$C_A(X) = \{a \in A : ax = xa \text{ for every } x \in X\}.$$

It is easy to check that centralizers are always subalgebras.

The key idea in the next proof is that a subalgebra B of A makes A into a (B, A)-bimodule, and that the centralizer of B can be described in terms of an endomorphism ring (this idea is exploited in proofs of the Morita theorems).

Theorem 9.115 (Double Centralizer). *Let A be a central simple algebra over a field k and let B be a simple subalgebra of A.*

(i) *$C_A(B)$ is a simple k-algebra.*

(ii) *$B \otimes_k A^{\mathrm{op}} \cong \mathrm{Mat}_s(\Delta)$ and $C_A(B) \cong \mathrm{Mat}_r(\Delta)$ for some division algebra Δ, where $r \mid s$.*

(iii) *$[B : k][C_A(B) : k] = [A : k]$.*

(iv) *$C_A(C_A(B)) = B$.*

Proof. Associativity of the multiplication in A shows that A can be viewed as a (B, A)-bimodule. As such, it is a left $(B \otimes_k A^{\mathrm{op}})$-module, where $(b \otimes a)x = bxa$ for all $x \in A$; we denote this module by A^*. But $B \otimes_k A^{\mathrm{op}}$ is a simple k-algebra, by Theorem 9.112, so that Corollary 8.63 gives $B \otimes_k A^{\mathrm{op}} \cong \mathrm{Mat}_s(\Delta)$ for some integer s and some division algebra Δ over k; in fact, $B \otimes_k A^{\mathrm{op}}$ has a unique (to isomorphism) minimal left ideal L, and $\Delta^{\mathrm{op}} \cong \mathrm{End}_{B \otimes_k A^{\mathrm{op}}}(L)$. Therefore, as $(B \otimes_k A^{\mathrm{op}})$-modules, Corollary 8.44 gives $A^* \cong L^r$, the direct sum of r copies of L, and so $\mathrm{End}_{B \otimes_k A^{\mathrm{op}}}(A^*) \cong \mathrm{Mat}_r(\Delta)$.

We claim that

$$C_A(B) \cong \mathrm{End}_{B \otimes_k A^{\mathrm{op}}}(A^*) \cong \mathrm{Mat}_r(\Delta);$$

this will prove (i) and most of (ii). If $\varphi \in \mathrm{End}_{B \otimes_k A^{\mathrm{op}}}(A^*)$, then it is, in particular, an endomorphism of A as a right A-module. Hence, for all $a \in A$, we have

$$\varphi(a) = \varphi(1a) = \varphi(1)a = ua,$$

where $u = \varphi(1)$. In particular, if $b \in B$, then $\varphi(b) = ub$. On the other hand, taking the left action of B into account, we have $\varphi(b) = \varphi(b1) = b\varphi(1) = bu$. Therefore, $ub = bu$ for all $b \in B$, and so $u \in C_A(B)$. Thus, $\varphi \mapsto \varphi(1)$ is a function $\mathrm{End}_{B \otimes_k A^{\mathrm{op}}}(A^*) \to C_A(B)$. It is routine to check that this function is an injective k-algebra map; it is also surjective, for if $u \in C_A(B)$, then the map $A \to A$, defined by $a \mapsto ua$, is a $(B \otimes_k A^{\mathrm{op}})$-map.

We now compute dimensions. Define $d = [\Delta : k]$. Since L is a minimal left ideal in $\text{Mat}_s(\Delta)$, we have $\text{Mat}_s(\Delta) \cong L^s$ (concretely, $L = \text{COL}(1)$, consisting of all first columns of $s \times s$ matrices over Δ). Therefore, $[\text{Mat}_s(\Delta) : k] = s^2[\Delta : k]$ and $[L^s : k] = s[L : k]$, so that

$$[L : k] = sd.$$

Also,

$$[A : k] = [A^* : k] = [L^r : k] = rsd.$$

It follows that

$$[A : k][B : k] = [B \otimes_k A^{\text{op}} : k] = [\text{Mat}_s(\Delta) : k] = s^2 d.$$

Therefore, $[B : k] = \frac{s^2 d}{rsd} = \frac{s}{r}$, and so $r \mid s$. Hence,

$$[B : k][C_A(B) : k] = [B : k][\text{Mat}_r(\Delta) : k] = \frac{s}{r} \cdot r^2 d = rsd = [A : k],$$

because we have already proved that $C_A(B) \cong \text{Mat}_r(\Delta)$.

Finally, we prove (iv). It is easy to see that $B \subseteq C_A(C_A(B))$: after all, if $b \in B$ and $u \in C_A(B)$, then $bu = ub$, and so b commutes with every such u. But $C_A(B)$ is a simple subalgebra, by (i), and so the equation in (iii) holds if we replace B by $C_A(B)$:

$$[C_A(B) : k][C_A(C_A(B)) : k] = [A : k].$$

We conclude that $[B : k] = [C_A(C_A(B)) : k]$; together with $B \subseteq C_A(C_A(B))$, this equality gives $B = C_A(C_A(B))$. •

Here is a minor variant of the theorem.

Corollary 9.116. *If B is a simple subalgebra of a central simple k-algebra A, where k is a field, then there is a division algebra D with $B^{\text{op}} \otimes_k A \cong \text{Mat}_s(D)$.*

Proof. By Theorem 9.115(ii), we have $B \otimes_k A^{\text{op}} \cong \text{Mat}_s(\Delta)$ for some division algebra Δ. Hence, $(B \otimes_k A^{\text{op}})^{\text{op}} \cong (\text{Mat}_s(\Delta))^{\text{op}}$. But $(\text{Mat}_s(\Delta))^{\text{op}} \cong \text{Mat}_s(\Delta^{\text{op}})$, by Proposition 8.13, while $(B \otimes_k A^{\text{op}})^{\text{op}} \cong B^{\text{op}} \otimes_k A$, by Exercise 9.65 on page 726. Setting $D = \Delta^{\text{op}}$ completes the proof. •

If Δ is a division algebra over a field k and if $\delta \in \Delta$, then the subdivision algebra generated by k and δ is a field, because elements in the center k commute with δ. We are interested in maximal subfields of Δ.

Lemma 9.117. *If Δ is a division algebra over a field k, then a subfield E of Δ is a maximal subfield if and only if $C_\Delta(E) = E$.*

Proof. If E is a maximal subfield of Δ, then $E \subseteq C_\Delta(E)$ because E is commutative. For the reverse inclusion, it is easy to see that if $\delta \in C_\Delta(E)$, then the division algebra E' generated by E and δ is a field. Hence, if $\delta \notin E$, then $E \subsetneq E'$, and the maximality of E is contradicted.

Conversely, suppose that E is a subfield with $C_\Delta(E) = E$. If E is not a maximal subfield of Δ, then there exists a subfield E' with $E \subsetneq E'$. Now $E' \subseteq C_\Delta(E)$, so that if there is some $a' \in E'$ with $a' \notin E$, then $E \neq C_\Delta(E)$. Therefore, E is a maximal subfield. •

After proving an elementary lemma about tensor products, we will extend the next result from division algebras to central simple algebras (see Theorem 9.127).

Theorem 9.118. *If D is a division algebra over a field k and E is a maximal subfield of D, then E is a splitting field for D; that is, $E \otimes_k D \cong \text{Mat}_s(E)$, where $s = [D : E] = [E : k]$.*

Proof. Let us specialize the algebras in Theorem 9.115. Here, $A = D$, $B = E$, and $C_A(E) = E$, by Lemma 9.117. Now the condition $C_A(B) \cong \text{Mat}_r(\Delta)$ becomes $E \cong \text{Mat}_r(\Delta)$; since E is commutative, $r = 1$ and $\Delta = E$. Thus, Corollary 9.116 says that $E \otimes_k D = E^{\text{op}} \otimes_k D \cong \text{Mat}_s(E)$.

The equality in Theorem 9.115(iii) is now $[D : k] = [E : k][E : k] = [E : k]^2$. But $[E \otimes_k D : k] = [\text{Mat}_s(E) : k] = s^2[E : k]$, so that $s^2 = [D : k] = [E : k]^2$ and $s = [E : k]$. •

Corollary 9.119. *If D is a division algebra over a field k, then all maximal subfields have the same degree over k.*

Proof. For every maximal subfield E, we have $[E : k] = [D : E] = \sqrt{[D : k]}$. •

This corollary can be illustrated by Example 9.113. The quaternions $\mathbb{H}$ is a four-dimensional $\mathbb{R}$-algebra, and so a maximal subfield must have degree 2 over $\mathbb{R}$. And so it is, for $\mathbb{C}$ is a maximal subfield.

We now prove a technical theorem that will yield wonderful results. Recall that a *unit* in a noncommutative ring A is an element having a two-sided inverse in A.

Theorem 9.120. *Let k be a field, let B be a simple k-algebra, and let A be a central simple k-algebra. If there are algebra maps $f, g \colon B \to A$, then there exists a unit $u \in A$ with*

$$g(b) = uf(b)u^{-1}$$

for all $b \in B$.

Proof. The map f makes A into a left B-module if we define the action of $b \in B$ on an element $a \in A$ as $f(b)a$. This action makes A into a (B, A)-bimodule, for the associative law in A gives $\big(f(b)x\big)a = f(b)(xa)$ for all $x \in A$. As usual, this (B, A)-bimodule is a left $(B \otimes_k A^{\text{op}})$-module, where $(b \otimes a')a = baa'$ for all $a \in A$; denote it by ${}_fA$. Similarly, g can be used to make A into a left $(B \otimes_k A^{\text{op}})$-module we denote by ${}_gA$. By Theorem 9.112, $B \otimes_k A^{\text{op}}$ is a simple k-algebra. Now

$$[{}_fA : \Delta] = [A : \Delta] = [{}_gA : \Delta],$$

so that ${}_fA \cong {}_gA$ as $(B \otimes_k A^{\mathrm{op}})$-modules, by Corollary 8.63. If $\varphi\colon {}_fA \to {}_gA$ is a $(B \otimes_k A^{\mathrm{op}})$-isomorphism, then

$$\varphi(f(b)aa') = g(b)\varphi(a)a' \tag{4}$$

for all $b \in B$ and $a, a' \in A$. Since φ is an automorphism of A as a right module over itself, $\varphi(a) = \varphi(1a) = ua$, where $u = \varphi(1) \in A$. To see that u is a unit, note that $\varphi^{-1}(a) = u'a$ for all $a \in A$. Now $a = \varphi\varphi^{-1}(a) = \varphi(u'a) = uu'a$ for all $a \in A$; in particular, when $a = 1$, we have $1 = uu'$. The equation $\varphi^{-1}\varphi = 1_A$ gives $1 = u'u$, as desired. Substituting into Eq. (4), we have

$$uf(b)a = \varphi(f(b)a) = g(b)\varphi(a) = g(b)ua$$

for all $a \in A$. In particular, if $a = 1$, then $uf(b) = g(b)u$ and $g(b) = uf(b)u^{-1}$. •

Corollary 9.121 (Skolem–Noether). *Let A be a central simple k-algebra over a field k, and let B and B' be isomorphic simple k-subalgebras of A. If $\psi\colon B \to B'$ is an isomorphism, then there exists a unit $u \in A$ with $\psi(b) = ubu^{-1}$ for all $b \in B$.*

Proof. In the theorem, take $f\colon B \to A$ to be the inclusion, define $B' = \operatorname{im}\psi$, and define $g = i\psi$, where $i\colon B' \to A$ is the inclusion. •

There is an analog of the Skolem–Noether theorem in group theory. A theorem of G. Higman, B. H. Neumann, and H. Neumann says that if B and B' are isomorphic subgroups of a group G, say, $\varphi\colon B \to B'$ is an isomorphism, then there exists a group G^* containing G and an element $u \in G^*$ with $\varphi(b) = ubu^{-1}$ for every $b \in B$. There is a proof in Rotman, *An Introduction to the Theory of Groups*, page 404.

Corollary 9.122. *Let k be a field. If ψ is an automorphism of* $\mathrm{Mat}_n(k)$, *then there exists a nonsingular matrix $P \in \mathrm{Mat}_n(k)$ with*

$$\psi(T) = PTP^{-1}$$

for every matrix T in $\mathrm{Mat}_n(k)$.

Proof. The matrix ring $A = \mathrm{Mat}_n(k)$ is a central simple k-algebra. Set $B = B' = A$ in the Skolem–Noether theorem. •

The following proof of Wedderburn's theorem is due to B. L. van der Waerden.

Theorem 9.123 (Wedderburn). *Every finite division ring D is a field.*

Proof. Let $Z = Z(D)$, and let E be a maximal subfield of D. If $d \in D$, then $Z(d)$ is a subfield of D, and hence there is a maximal subfield E_d containing $Z(d)$. By Corollary 9.119, all maximal subfields have the same degree, hence have the same order. By Corollary 3.132, all maximal subfields here are isomorphic.[19] For every $d \in D$, the

[19] It is not true that maximal subfields in arbitrary division algebras are isomorphic; see Exercise 9.80.

Skolem–Noether theorem says there is $x_d \in D$ with $E_d = x_d E x_d^{-1}$. Therefore, $D = \bigcup_x xEx^{-1}$, and so

$$D^\times = \bigcup_x xE^\times x^{-1}.$$

If E is a proper subfield of D, then $E^\times$ is a proper subgroup of $D^\times$, and this equation contradicts Exercise 5.32 on page 278. Therefore, $D = E$ is commutative. •

Theorem 9.124 (Frobenius). *If D is a noncommutative finite-dimensional real division algebra, then $D \cong \mathbb{H}$.*

Proof. If E is a maximal subfield of D, then $[D : E] = [E : \mathbb{R}] \le 2$. If $[E : \mathbb{R}] = 1$, then $[D : \mathbb{R}] = 1^2 = 1$ and $D = \mathbb{R}$. Hence, $[E : \mathbb{R}] = 2$ and $[D : \mathbb{R}] = 4$. Let us identify E with $\mathbb{C}$ (we know they are isomorphic). Now complex conjugation is an automorphism of E, so that the Skolem–Noether theorem gives $x \in D$ with $\overline{z} = xzx^{-1}$ for all $z \in E$. In particular, $-i = xix^{-1}$. Hence,

$$x^2ix^{-2} = x(-i)x^{-1} = -xix^{-1} = i,$$

and so x^2 commutes with i. Therefore, $x^2 \in C_D(E) = E$, by Lemma 9.117, and so $x^2 = a + bi$ for $a, b \in \mathbb{R}$. But

$$a + bi = x^2 = xx^2x^{-1} = x(a + bi)x^{-1} = a - bi,$$

so that $b = 0$ and $x^2 \in \mathbb{R}$. If $x^2 > 0$, then there is $t \in \mathbb{R}$ with $x^2 = t^2$. Now $(x+t)(x-t) = 0$ gives $x = \pm t \in \mathbb{R}$, contradicting $-i = xix^{-1}$. Therefore, $x^2 = -r^2$ for some real r. The element j, defined by $j = x/r$, satisfies $j^2 = -1$ and $ji = -ij$. The list $1, i, j, ij$ is linearly independent over $\mathbb{R}$: if $a+bi+cj+dij = 0$, then $(-di-c)j = a+ib \in \mathbb{C}$. Since $j \notin \mathbb{C}$ (lest $x \in \mathbb{C}$), we must have $-di - c = 0 = a + bi$. Hence, $a = b = 0 = c = d$. Since $[D : \mathbb{R}] = 4$, the list $1, i, j, ij$ is a basis of D. It is now routine to see that if we define $k = ij$, then $ki = j = -ik$, $jk = i = -kj$, and $k^2 = -1$, and so $D \cong \mathbb{H}$. •

In 1929, R. Brauer introduced the Brauer group to study division rings. Since construction of division rings was notoriously difficult, he considered the wider class of central simple algebras. Brauer introduced the following relation on central simple k-algebras.

Definition. Two central simple k-algebras A and B are ***similar***, denoted by $A \sim B$, if there are integers n and m with

$$A \otimes_k \mathrm{Mat}_n(k) \cong B \otimes_k \mathrm{Mat}_m(k).$$

By the Wedderburn theorem, $A \cong \mathrm{Mat}_n(\Delta)$ for a unique division algebra Δ over k, and we shall see that $A \sim B$ if and only if they determine the same division algebra.

Lemma 9.125. *Let A be a finite-dimensional algebra over a field k. If S and T are k-subalgebras of A such that*

(i) $st = ts$ *for all* $s \in S$ *and* $t \in T$;

(ii) $A = ST$;

(iii) $[A : k] = [S : k][T : k]$,

then $A \cong S \otimes_k T$.

Proof. There is a k-linear transformation $f \colon S \otimes_k T \to A$ with $s \otimes t \mapsto st$, because $(s, t) \mapsto st$ is a k-bilinear function $S \times T \to A$. Condition (i) implies that f is an algebra map, for

$$f\big((s \otimes t)(s' \otimes t')\big) = f(ss' \otimes tt') = ss'tt' = sts't' = f(s \otimes t) f(s' \otimes t').$$

Since $A = ST$, by condition (ii), the k-linear transformation f is a surjection; since $\dim_k(S \otimes_k T) = \dim_k(A)$, by condition (iii), f is a k-algebra isomorphism. •

Lemma 9.126. *Let k be a field.*

(i) *If A is a k-algebra, then*

$$A \otimes_k \mathrm{Mat}_n(k) \cong \mathrm{Mat}_n(A).$$

(ii) $\mathrm{Mat}_n(k) \otimes_k \mathrm{Mat}_m(k) \cong \mathrm{Mat}_{nm}(k)$.

(iii) *$A \sim B$ is an equivalence relation.*

(iv) *If A is a central simple algebra, then*

$$A \otimes_k A^{\mathrm{op}} \cong \mathrm{Mat}_n(k),$$

where $n = [A : k]$.

Proof. (i) Define k-subalgebras of $\mathrm{Mat}_n(A)$ by

$$S = \mathrm{Mat}_n(k) \quad \text{and} \quad T = \{aI : a \in A\}.$$

If $s \in S$ and $t \in T$, then $st = ts$ (for the entries of matrices in S commute with elements $a \in A$). Now S contains every matrix unit E_{ij} (whose ij entry is 1 and whose other entries are 0), so that ST contains all matrices of the form $a_{ij}E_{ij}$ for all ij, where $a_{ij} \in A$; hence, $ST = \mathrm{Mat}_n(A)$. Finally, $[S : k][T : k] = n^2[A : k] = [\mathrm{Mat}_n(A) : k]$. Therefore, Lemma 9.125 gives the desired isomorphism.

(ii) If V and W are vector spaces over k of dimensions n and m, respectively, it suffices to prove that $\mathrm{End}_k(V) \otimes_k \mathrm{End}_k(W) \cong \mathrm{End}_k(V \otimes_k W)$. Define S to be all $f \otimes 1_W$, where $f \in \mathrm{End}_k(V)$, and define T to be all $1_V \otimes g$, where $g \in \mathrm{End}_k(W)$. It is routine to check that the three conditions in Lemma 9.125 hold.

(iii) Since $k = \mathrm{Mat}_1(k)$, we have $A \cong A \otimes_k k \cong A \otimes_k \mathrm{Mat}_1(k)$, so that $\sim$ is reflexive. Symmetry is obvious; for transitivity, suppose that $A \sim B$ and $B \sim C$; that is,

$$A \otimes_k \mathrm{Mat}_n(k) \cong B \otimes_k \mathrm{Mat}_m(k) \quad \text{and} \quad B \otimes_k \mathrm{Mat}_r(k) \cong C \otimes_k \mathrm{Mat}_s(k).$$

Then $A \otimes_k \mathrm{Mat}_n(k) \otimes_k \mathrm{Mat}_r(k) \cong A \otimes_k \mathrm{Mat}_{nr}(A)$, by part (ii). On the other hand,

$$\begin{aligned} A \otimes_k \mathrm{Mat}_n(k) \otimes_k \mathrm{Mat}_r(k) &\cong B \otimes_k \mathrm{Mat}_m(k) \otimes_k \mathrm{Mat}_r(k) \\ &\cong C \otimes_k \mathrm{Mat}_m(k) \otimes_k \mathrm{Mat}_s(k) \\ &\cong C \otimes_k \mathrm{Mat}_{ms}(k). \end{aligned}$$

Therefore, $A \sim C$, and so $\sim$ is an equivalence relation.

(iv) Define $f\colon A \times A^{\mathrm{op}} \to \mathrm{End}_k(A)$ by $f(a, c) = \lambda_a \circ \rho_c$, where $\lambda_a\colon x \mapsto ax$ and $\rho_c\colon x \mapsto xc$; it is routine to check that λ_a and ρ_c are k-maps (so their composite is also a k-map), and that f is k-biadditive. Hence, there is a k-map $\widehat{f}\colon A \otimes_k A^{\mathrm{op}} \to \mathrm{End}_k(A)$ with $\widehat{f}(a \otimes c) = \lambda_a \circ \rho_c$. Associativity $a(xc) = (ax)c$ in A says that $\lambda_a \circ \rho_c = \rho_c \circ \lambda_a$, from which it easily follows that $\widehat{f}$ is a k-algebra map. As $A \otimes_k A^{\mathrm{op}}$ is a simple k-algebra and $\ker \widehat{f}$ is a proper two-sided ideal, we have $\widehat{f}$ injective. Now $\dim_k(\mathrm{End}_k(A)) = \dim_k(\mathrm{Hom}_k(A, A)) = n^2$, where $n = [A : k]$. Since $\dim_k(\mathrm{im}\, \widehat{f}) = \dim_k(A \otimes_k A^{\mathrm{op}}) = n^2$, it follows that $\widehat{f}$ is a k-algebra isomorphism: $A \otimes_k A^{\mathrm{op}} \cong \mathrm{End}_k(A)$. •

We now extend Theorem 9.118 from division algebras to central simple algebras.

Theorem 9.127. *Let A be a central simple k-algebra over a field k, so that $A \cong \mathrm{Mat}_r(\Delta)$, where Δ is a division algebra over k. If E is a maximal subfield of Δ, then E splits A; that is, there is an integer n and an isomorphism*

$$E \otimes_k A \cong \mathrm{Mat}_n(E).$$

More precisely, if $[\Delta : E] = s$, then $n = rs$ and $[A : k] = (rs)^2$.

Proof. By Theorem 9.118, Δ is split by a maximal subfield E (which is, of course, a finite extension of k): $E \otimes_k \Delta \cong \mathrm{Mat}_s(E)$, where $s = [\Delta : E] = [E : k]$. Hence,

$$\begin{aligned} E \otimes_k A &\cong E \otimes_k \mathrm{Mat}_r(\Delta) \cong E \otimes_k (\Delta \otimes_k \mathrm{Mat}_r(k)) \\ &\cong (E \otimes_k \Delta) \otimes_k \mathrm{Mat}_r(k) \cong \mathrm{Mat}_s(E) \otimes_k \mathrm{Mat}_r(k) \cong \mathrm{Mat}_{rs}(E). \end{aligned}$$

Therefore, $A \cong \mathrm{Mat}_r(\Delta)$ gives $[A : k] = r^2[\Delta : k] = r^2 s^2$. •

Definition. If $[A]$ denotes the equivalence class of a central simple k-algebra A under similarity, define the ***Brauer group*** $\mathrm{Br}(k)$ to be the set

$$\mathrm{Br}(k) = \big\{[A] : A \text{ is a central simple } k\text{-algebra}\big\}$$

with binary operation

$$[A][B] = [A \otimes_k B].$$

Theorem 9.128. $\mathrm{Br}(k)$ *is an abelian group for every field* k*. Moreover, if* $A \cong \mathrm{Mat}_n(\Delta)$ *for a division algebra* Δ*, then* Δ *is central simple and* $[A] = [\Delta]$ *in* $\mathrm{Br}(k)$*.*

Proof. We show that the operation is well-defined: If A, A', B, B' are k-algebras with $A \sim A'$ and $B \sim B'$, then $A \otimes_k B \sim A' \otimes_k B'$. The isomorphisms

$$A \otimes_k \mathrm{Mat}_n(k) \cong A' \otimes_k \mathrm{Mat}_m(k) \quad \text{and} \quad B \otimes_k \mathrm{Mat}_r(k) \cong B' \otimes_k \mathrm{Mat}_s(k)$$

give $A \otimes_k B \otimes_k \mathrm{Mat}_n(k) \otimes_k \mathrm{Mat}_r(k) \cong A' \otimes_k B' \otimes_k \mathrm{Mat}_m(k) \otimes_k \mathrm{Mat}_s(k)$ (we are using commutativity and associativity of tensor product), so that Lemma 9.126(ii) gives $A \otimes_k B \otimes_k \mathrm{Mat}_{nr}(k) \cong A' \otimes_k B' \otimes_k \mathrm{Mat}_{ms}(k)$. Therefore, $A \otimes_k B \sim A' \otimes_k B'$.

That $[k]$ is the identity follows from $k \otimes_k A \cong A$, associativity and commutativity follow from associativity and commutativity of tensor product, and Lemma 9.126(iv) shows that $[A]^{-1} = [A^{\mathrm{op}}]$. Therefore, $\mathrm{Br}(k)$ is an abelian group.

If A is a central simple k-algebra, then $A \cong \mathrm{Mat}_r(\Delta)$ for some finite-dimensional division algebra Δ over k. Hence, $k = Z(A) \cong Z(\mathrm{Mat}_r(\Delta)) \cong Z(\Delta)$, by Theorem 9.112. Thus, Δ is a central simple k-algebra, $[\Delta] \in \mathrm{Br}(k)$, and $[\Delta] = [A]$ (because $\Delta \otimes_k \mathrm{Mat}_r(k) \cong \mathrm{Mat}_r(\Delta) \cong A \cong A \otimes_k k \cong A \otimes_k \mathrm{Mat}_1(k)$). •

The next proposition shows the significance of the Brauer group.

Proposition 9.129. *If* k *is a field, then there is a bijection from* $\mathrm{Br}(k)$ *to the family* $\mathcal{D}$ *of all isomorphism classes of finite-dimensional division algebras over* k*, and so* $|\mathrm{Br}(k)| = |\mathcal{D}|$*. Therefore, there exists a noncommutative division ring, finite-dimensional over its center* k*, if and only if* $\mathrm{Br}(k) \neq \{0\}$*.*

Proof. Define a function $\varphi \colon \mathrm{Br}(k) \to \mathcal{D}$ by setting $\varphi([A])$ to be the isomorphism class of Δ if $A \cong \mathrm{Mat}_n(\Delta)$. Note that Theorem 9.128 shows that $[A] = [\Delta]$ in $\mathrm{Br}(k)$. Let us see that φ is well-defined. If $[\Delta] = [\Delta']$, then $\Delta \sim \Delta'$, so there are integers n and m with $\Delta \otimes_k \mathrm{Mat}_n(k) \cong \Delta' \otimes_k \mathrm{Mat}_m(k)$. Hence, $\mathrm{Mat}_n(\Delta) \cong \mathrm{Mat}_m(\Delta')$. By the uniqueness in the Wedderburn–Artin theorems, $\Delta \cong \Delta'$ (and $n = m$). Therefore, $\varphi([\Delta]) = \varphi([\Delta'])$.

Clearly, φ is surjective, for if Δ is a finite-dimensional division algebra over k, then the isomorphism class of Δ is equal to $\varphi([\Delta])$. To see that φ is injective, suppose that $\varphi([\Delta]) = \varphi([\Delta'])$. Then, $\Delta \cong \Delta'$, which implies $\Delta \sim \Delta'$. •

Example 9.130.

(i) If k is an algebraically closed field, then Theorem 9.114 shows that $\mathrm{Br}(k) = \{0\}$.

(ii) If k is a finite field, then Wedderburn's Theorem 9.123 (= Theorem 8.23) shows that $\mathrm{Br}(k) = \{0\}$.

(iii) If $k = \mathbb{R}$, then Frobenius's Theorem 9.124 shows that $\mathrm{Br}(\mathbb{R}) \cong \mathbb{I}_2$.

(iv) It is proved, using class field theory, that $\mathrm{Br}(\mathbb{Q}_p) \cong \mathbb{Q}/\mathbb{Z}$, where $\mathbb{Q}_p$ is the field of p-adic numbers. Moreover, there is an exact sequence

$$0 \to \mathrm{Br}(\mathbb{Q}) \to \mathrm{Br}(\mathbb{R}) \oplus \sum_p \mathrm{Br}(\mathbb{Q}_p) \xrightarrow{\varphi} \mathbb{Q}/\mathbb{Z} \to 0.$$

If we write $\text{Br}(\mathbb{R}) = \langle \frac{1}{2} + \mathbb{Z} \rangle \subseteq \mathbb{Q}/\mathbb{Z}$, then φ is the "sum of coordinates" map.

In a series of deep papers, $\text{Br}(k)$ was computed for the most interesting fields k arising in algebraic number theory (*local fields*, one of which is $\mathbb{Q}_p$, and *global fields*) by A. A. Albert, R. Brauer, H. Hasse, and E. Noether. ◂

Proposition 9.131. *If E/k is a field extension, then there is a homomorphism*

$$f_{E/k} \colon \text{Br}(k) \to \text{Br}(E)$$

given by $[A] \mapsto [E \otimes_k A]$.

Proof. If A and B are central simple k-algebras, then $E \otimes_k A$ and $E \otimes_k B$ are central simple E-algebras, by Theorem 9.112. If $A \sim B$, then $E \otimes_k A \sim E \otimes_k B$ as E-algebras, by Exercise 9.77 on page 740. It follows that the function $f_{E/k}$ is well-defined. Finally, $f_{E/k}$ is a homomorphism, because

$$(E \otimes_k A) \otimes_E (E \otimes_k B) \cong (E \otimes_E E) \otimes_k (A \otimes_k B) \cong E \otimes_k (A \otimes_k B),$$

by Proposition 8.84, associativity of tensor product. •

Definition. If E/k is a field extension, then the ***relative Brauer group***, $\text{Br}(E/k)$, is the kernel of homomorphism $f_{E/k} \colon \text{Br}(k) \to \text{Br}(E)$:

$$\text{Br}(E/k) = \ker f_{E/k} = \big\{[A] \in \text{Br}(k) : A \text{ is split by } E\big\}.$$

Corollary 9.132. *For every field k, we have*

$$\text{Br}(k) = \bigcup_{E/k \text{ finite}} \text{Br}(E/k).$$

Proof. This follows at once from Theorem 9.127. •

In a word, the Brauer group arose as a way to study division rings. It is an interesting object, but we have not really used it seriously. For example, we still know no noncommutative division rings other than the real division algebra $\mathbb{H}$ and its variants for subfields k of $\mathbb{R}$. We will remedy this when we introduce *crossed product algebras* in Chapter 10. For example, we will see, in Corollary 10.133, that there exists a division ring whose center is a field of characteristic $p > 0$. For further developments, we refer the reader to Jacobson, *Finite-Dimensional Division Algebras over Fields*, and Reiner, *Maximal Orders*.

EXERCISES

9.72 Prove that $\mathbb{H} \otimes_{\mathbb{R}} \mathbb{H} \cong \mathrm{Mat}_4(\mathbb{R})$ as $\mathbb{R}$-algebras.
Hint. Use Corollary 8.60 for the central simple $\mathbb{R}$-algebra $\mathbb{H} \otimes_{\mathbb{R}} \mathbb{H}$.

9.73 We have given one isomorphism $\mathbb{C} \otimes_{\mathbb{R}} \mathbb{H} \cong \mathrm{Mat}_2(\mathbb{C})$ in Example 9.113. Describe all possible isomorphisms between these two algebras.
Hint. Use the Skolem–Noether theorem.

9.74 Prove that $\mathbb{C} \otimes_{\mathbb{R}} \mathbb{C} \cong \mathbb{C} \times \mathbb{C}$ as $\mathbb{R}$-algebras.

9.75 (i) Let $\mathbb{C}(x)$ and $\mathbb{C}(y)$ be function fields. Prove that $R = \mathbb{C}(x) \otimes_{\mathbb{C}} \mathbb{C}(y)$ is isomorphic to a subring of $\mathbb{C}(x, y)$. Conclude that R has no zero divisors.

(ii) Prove that $\mathbb{C}(x) \otimes_{\mathbb{C}} \mathbb{C}(y)$ is not a field.

Hint. Show that R is isomorphic to the subring of $\mathbb{C}(x, y)$ consisting of polynomials of the form $f(x, y)/g(x)h(y)$.

(iii) Use Exercise 8.39 on page 573 to prove that the tensor product of artinian algebras need not be artinian.

9.76 Let A be a central simple k-algebra. If A is split by a field E, prove that A is split by any field extension E' of E.

9.77 Let E/k be a field extension. If A and B are central simple k-algebras with $A \sim B$, prove that $E \otimes_k A \sim E \otimes_k B$ as central simple E-algebras.

9.78 If D is a finite-dimensional division algebra over $\mathbb{R}$, prove that D is isomorphic to either $\mathbb{R}$, $\mathbb{C}$, or $\mathbb{H}$.

9.79 Prove that $\mathrm{Mat}_2(\mathbb{H}) \cong \mathbb{H} \otimes_{\mathbb{R}} \mathrm{Mat}_2(\mathbb{R})$ as $\mathbb{R}$-algebras.

9.80 (i) Let A be a four-dimensional vector space over $\mathbb{Q}$, and let $1, i, j, k$ be a basis. Show that A is a division algebra if we define 1 to be the identity and

$$\begin{array}{lll} i^2 = -1 & j^2 = -2 & k^2 = -2 \\ ij = \ k & jk = \ 2i & ki = \ j \\ ji = -k & kj = -2i & ik = -j \end{array}$$

Prove that A is a division algebra over $\mathbb{Q}$.

(ii) Prove that $\mathbb{Q}(i)$ and $\mathbb{Q}(j)$ are nonisomorphic maximal subfields of A.

9.81 Let D be the $\mathbb{Q}$-subalgebra of $\mathbb{H}$ having basis $1, i, j, k$.

(i) Prove that D is a division algebra over $\mathbb{Q}$.

Hint. Compute the center $Z(D)$.

(ii) For any pair of nonzero rationals p and q, prove that D has a maximal subfield isomorphic to $\mathbb{Q}(\sqrt{-p^2 - q^2})$.

Hint. Compute $(pi + qj)^2$.

9.82 (**Dickson**) If D is a division algebra over a field k, then each $d \in D$ is algebraic over k. Prove that $d, d' \in D$ are conjugate in D if and only if $\mathrm{irr}(d, k) = \mathrm{irr}(d', k)$.
Hint. Use the Skolem–Noether theorem.

9.83 Prove that if A is a central simple k-algebra with $A \sim \mathrm{Mat}_n(k)$, then $A \cong \mathrm{Mat}_m(k)$ for some integer m.

9.84 Prove that if A is a central simple k-algebra with $[A]$ of finite order m in $\mathrm{Br}(k)$, then

$$A \otimes_k \cdots \otimes_k A \cong \mathrm{Mat}_r(k)$$

(there are m factors equal to A) for some integer r. (In Chapter 10, we shall see that every element in $\mathrm{Br}(k)$ has finite order.)

9.8 EXTERIOR ALGEBRA

In calculus, the ***differential*** df of a differentiable function $f(x, y)$ at a point $P = (x_0, y_0)$ is defined by

$$df|_P = \tfrac{\partial f}{\partial x}|_P(x - x_0) + \tfrac{\partial f}{\partial y}|_P(y - y_0).$$

If (x, y) is a point near P, then $df|_P$ approximates the difference between the true value $f(x, y)$ and $f(x_0, y_0)$. The quantity df is considered "small," and so its square, a second-order approximation, is regarded as negligible. For the moment, let us take being negligible seriously: Suppose that

$$(df)^2 \approx 0$$

for all differentials df. There is a curious consequence: if du and dv are differentials, then so is $du + dv = d(u + v)$. But $(du + dv)^2 \approx 0$ gives

$$\begin{aligned} 0 &\approx (du + dv)^2 \\ &\approx (du)^2 + du\,dv + dv\,du + (dv)^2 \\ &\approx du\,dv + dv\,du, \end{aligned}$$

and so du and dv anticommute:

$$dv\,du \approx -du\,dv.$$

Now consider a double integral $\iint_D f(x, y)dx\,dy$, where D is some region in the plane. Equations

$$\begin{aligned} x &= F(u, v) \\ y &= G(u, v) \end{aligned}$$

lead to the change of variables formula:

$$\iint_D f(x, y)dx\,dy = \iint_\Delta f(F(u, v), G(u, v))J\,du\,dv,$$

where Δ is some new region and J is the ***Jacobian***:

$$J = \left|\det \begin{bmatrix} F_u & F_v \\ G_u & G_v \end{bmatrix}\right|.$$

A key idea in the proof of this formula is that the graph of a differentiable function $f(x, y)$ looks, locally, like a real vector space—its tangent plane. Let us denote a basis of the tangent plane at a point by dx, dy. If du, dv is another basis of this tangent plane, then the chain rule defines a linear transformation by the following linear equations:

$$\begin{aligned} dx &= F_u du + F_v dv \\ dy &= G_u du + G_v dv. \end{aligned}$$

The Jacobian J now arises in a natural way.

$$\begin{aligned} dx\,dy &= (F_u du + F_v dv)(G_u du + G_v dv) \\ &= F_u du G_u du + F_u du G_v dv + F_v dv G_u du + F_v dv G_v dv \\ &= F_u G_u (du)^2 + F_u G_v du\,dv + F_v G_u dv\,du + F_v G_v (dv)^2 \\ &\approx F_u G_v du\,dv + F_v G_u dv\,du \\ &\approx (F_u G_v - F_v G_u) du\,dv \\ &= \det \begin{bmatrix} F_u & F_v \\ G_u & G_v \end{bmatrix} du dv. \end{aligned}$$

Analytic considerations, involving orientation, force us to use the absolute value of the determinant when proving the change of variables formula.

In the preceding equations, we used the distributive and associative laws, together with anticommutativity; that is, we assumed that the differentials form a ring in which all squares are 0. The following construction puts this kind of reasoning on a firm basis.

Definition. If M is a k-module, where k is a commutative ring, then its ***exterior algebra***[20] is $\bigwedge(M) = T(M)/J$, pronounced "wedge M," where J is the two-sided ideal generated by all $m \otimes m$ with $m \in M$. The image of $m_1 \otimes \cdots \otimes m_p$ in $\bigwedge(M)$ is denoted by

$$m_1 \wedge \cdots \wedge m_p.$$

Notice that J is generated by homogeneous elements (of degree 2), and so it is a graded ideal, by Proposition 9.95. Hence, $\bigwedge(M)$ is a graded k-algebra,

$$\bigwedge(M) = k \oplus M \oplus \bigwedge^2(M) \oplus \bigwedge^3(M) \oplus \cdots,$$

where, for $p \geq 2$, we have $\bigwedge^p(M) = T^p(M)/J^p$ and $J^p = J \cap T^p(M)$. Finally, $\bigwedge(M)$ is generated, as a k-algebra, by $\bigwedge^1(M) = M$.

Definition. We call $\bigwedge^p(M)$ the ***p*th exterior power** of a k-module M.

[20] The original adjective in this context—the German *ausserer*, meaning "outer"—was introduced by Grassmann in 1844. Grassmann used it in contrast to *inner product*. The first usage of the translation *exterior* can be found in work of E. Cartan in 1945, who wrote that he was using terminology of Kaehler. The wedge notation seems to have been introduced by Bourbaki.

Lemma 9.133. *Let k be a commutative ring, and let M be a k-module.*

(i) *If $m, m' \in M$, then in $\bigwedge^2(M)$, we have*

$$m \wedge m' = -m' \wedge m.$$

(ii) *If $p \geq 2$ and $m_i = m_j$ for some $i \neq j$, then $m_1 \wedge \cdots \wedge m_p = 0$ in $\bigwedge^p(M)$.*

Proof. (i) Recall that $\bigwedge^2(M) = (M \otimes_k M)/J^2$, where $J^2 = J \cap (M \otimes_k M)$. If $m, m' \in M$, then

$$(m + m') \otimes (m + m') = m \otimes m + m \otimes m' + m' \otimes m + m' \otimes m'.$$

Therefore,

$$m \otimes m' + J^2 = -m' \otimes m + J^2,$$

because J^2 contains $(m + m') \otimes (m + m')$, $m \otimes m$, and $m' \otimes m'$. It follows that

$$m \wedge m' = -m' \wedge m$$

for all $m, m' \in M$.

(ii) As we saw in the proof of Proposition 9.95, $\bigwedge^p(M) = T^p(M)/J^p$, where $J^p = J \cap T^p(M)$ consists of all elements of degree p in the ideal J generated by all elements in $T^2(M)$ of the form $m \otimes m$. In more detail, J^p consists of all sums of homogeneous elements $\alpha \otimes m \otimes m \otimes \beta$, where $m \in M$, $\alpha \in T^q(M)$, $\beta \in T^r(M)$, and $q + r + 2 = p$; it follows that $m_1 \wedge \cdots \wedge m_p = 0$ if there are two equal *adjacent* factors, say, $m_i = m_{i+1}$. Since multiplication in $\bigwedge(M)$ is associative, however, we can (anti)commute a factor m_i of $m_1 \wedge \cdots \wedge m_p$ several steps away at the possible cost of a change in sign, and so we can force any pair of factors to be adjacent. •

One of our goals is to give a "basis-free" construction of determinants, and the idea is to focus on some properties that such a function has. If we regard an $n \times n$ matrix A as consisting of its n columns, then its determinant, $\det(A)$, is a function of n variables (each ranging over n-tuples). One property of determinants is that $\det(A) = 0$ if two columns of A are equal, and another property is that it is multilinear. It will be seen that these properties almost characterize the determinant.

Definition. If M and N are k-modules, a k-multilinear function $f \colon \times^p M \to N$ (where $\times^p M$ is the cartesian product of M with itself p times) is ***alternating*** if

$$f(m_1, \ldots, m_p) = 0$$

whenever $m_i = m_j$ for some $i \neq j$.

An alternating $\mathbb{R}$-bilinear function arises naturally when considering (signed) areas in the plane $\mathbb{R}^2$. If $v_1, v_2 \in \mathbb{R}^2$, define $A(v_1, v_2)$ to be the area of the parallelogram having sides v_1 and v_2. It is clear that

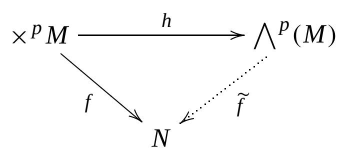
$$A(rv_1, sv_2) = rsA(v_1, v_2)$$

for all $r, s \in \mathbb{R}$ (but we must say what this means when these numbers are negative), and a geometric argument can be given to show that

$$A(w_1 + v_1, v_2) = A(w_1, v_2) + A(v_1, v_2);$$

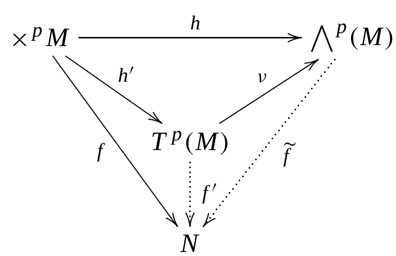
that is, A is $\mathbb{R}$-bilinear. Now A is alternating, for $A(v_1, v_1) = 0$ because the degenerate "parallelogram" having sides v_1 and v_1 has zero area. A similar argument shows that volume is an alternating $\mathbb{R}$-multilinear function on $\mathbb{R}^3$, as we see in vector calculus using the cross product.

Theorem 9.134. *For all $p \geq 0$ and all k-modules M, the pth exterior power $\bigwedge^p(M)$ solves the universal mapping problem posed by alternating multilinear functions.*



If $h\colon \times^p M \to \bigwedge^p(M)$ is defined by $h(m_1, \ldots, m_p) = m_1 \wedge \cdots \wedge m_p$, then for every alternating multilinear function f, there exists a unique k-homomorphism $\widetilde{f}$ making the diagram commute.

Proof. Consider the diagram



where $h'(m_1, \ldots, m_p) = m_1 \otimes \cdots \otimes m_p$ and $\nu(m_1 \otimes \cdots \otimes m_p) = m_1 \wedge \cdots \wedge m_p$. Since f is multilinear, there is a k-map $f'\colon T^p(M) \to N$ with $f'h' = f$; since f is alternating, $J \cap T^p(M) \subseteq \ker f'$, and so f' induces a map

$$\widetilde{f}\colon T^p(M)/(J \cap T^p(M)) \to N$$

with $\widetilde{f}\nu = f'$. Hence,

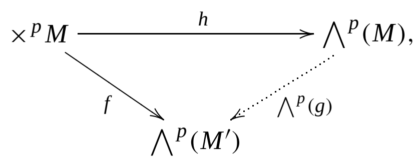
$$\widetilde{f}h = \widetilde{f}\nu h' = f'h' = f.$$

But $T^p(M)/(J \cap T^p(M)) = \bigwedge^p(M)$, as desired. Finally, $\widetilde{f}$ is the unique such map because $\operatorname{im} h$ generates $\bigwedge^p(M)$. •

Proposition 9.135. *For each $p \geq 0$, the pth exterior power is a functor*

$$\bigwedge^p : {}_k\mathbf{Mod} \to {}_k\mathbf{Mod}.$$

Proof. Now $\bigwedge^p(M)$ has been defined on modules; it remains to define it on morphisms. Suppose that $g \colon M \to M'$ is a k-homomorphism. Consider the diagram

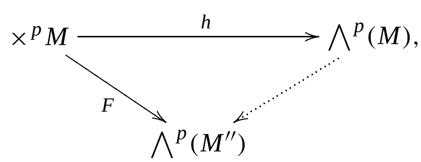


where $f(m_1, \ldots, m_p) = gm_1 \wedge \cdots \wedge gm_p$. It is easy to see that f is an alternating multilinear function, and so universality yields a unique map

$$\bigwedge^p(g) \colon \bigwedge^p(M) \to \bigwedge^p(M')$$

with $m_1 \wedge \cdots \wedge m_p \mapsto gm_1 \wedge \cdots \wedge gm_p$.

If g is the identity map on a module M, then $\bigwedge^p(g)$ is also the identity map, for it fixes a set of generators. Finally, suppose that $g' \colon M' \to M''$ is a k-map. It is routine to check that both $\bigwedge^p(g'g)$ and $\bigwedge^p(g')\bigwedge^p(g)$ make the following diagram commute



where $F(m_1, \ldots, m_p) = g'gm_1 \wedge \cdots \wedge g'gm_p$. Uniqueness of such a dashed arrow gives $\bigwedge^p(g'g) = \bigwedge^p(g')\bigwedge^p(g)$, as desired. •

We will soon see that $\bigwedge^p$ is not as nice as Hom or tensor, for it is not an additive functor.

Theorem 9.136 (Anticommutativity). *If M is a k-module, $x \in \bigwedge^p(M)$, and $y \in \bigwedge^q(M)$, then*

$$x \wedge y = (-1)^{pq} y \wedge x.$$

Remark. This identity holds only for products of homogeneous elements. ◂

Proof. If $x \in \bigwedge^0(M) = k$, then $\bigwedge(M)$ being a k-algebra implies $x \wedge y = y \wedge x$ for all $y \in \bigwedge(M)$, and so the identity holds, in particular, for $y \in \bigwedge^q(M)$ for any q. A similar argument holds if y is homogeneous of degree 0. Therefore, we may assume that $p, q \geq 1$; we do a double induction.

Base Step: $p = 1$ and $q = 1$. Suppose that $x, y \in \bigwedge^1(M) = M$. Now

$$\begin{aligned} 0 &= (x + y) \wedge (x + y) \\ &= x \wedge x + x \wedge y + y \wedge x + y \wedge y \\ &= x \wedge y + y \wedge x. \end{aligned}$$

It follows that $x \wedge y = -y \wedge x$, as desired.
Inductive step: $(p, 1) \Rightarrow (p + 1, 1)$. The inductive hypothesis gives

$$(x_1 \wedge \cdots \wedge x_p) \wedge y = (-1)^p y \wedge (x_1 \wedge \cdots \wedge x_p).$$

Using associativity, we have

$$\begin{aligned} (x_1 \wedge \cdots \wedge x_{p+1}) \wedge y &= x_1 \wedge [(x_2 \wedge \cdots \wedge x_{p+1}) \wedge y] \\ &= x_1 \wedge (-1)^p [y \wedge (x_2 \wedge \cdots \wedge x_{p+1})] \\ &= [x_1 \wedge (-1)^p y] \wedge (x_2 \wedge \cdots \wedge x_{p+1}) \\ &= (-1)^{p+1} (y \wedge x_1) \wedge (x_2 \wedge \cdots \wedge x_{p+1}). \end{aligned}$$

Inductive Step: $(p, q) \Rightarrow (p, q + 1)$. Assume that

$$(x_1 \wedge \cdots \wedge x_p) \wedge (y_1 \wedge \cdots \wedge y_q) = \\ (-1)^{pq} (y_1 \wedge \cdots \wedge y_q) \wedge (x_1 \wedge \cdots \wedge x_p).$$

We let the reader prove, using associativity, that

$$(x_1 \wedge \cdots \wedge x_p) \wedge (y_1 \wedge \cdots \wedge y_{q+1}) = \\ (-1)^{p(q+1)} (y_1 \wedge \cdots \wedge y_{q+1}) \wedge (x_1 \wedge \cdots \wedge x_p). \quad \bullet$$

Definition. Let n be a positive integer and let $1 \le p \le n$. An ***increasing*** $p \le n$***-list*** is a list

$$H = i_1, \ldots, i_p$$

for which $1 \le i_1 < i_2 < \cdots < i_p \le n$.

If $H = i_1, \ldots, i_p$ is an increasing $p \le n$-list, we write

$$e_H = e_{i_1} \wedge e_{i_2} \wedge \cdots \wedge e_{i_p}.$$

Of course, the number of increasing $p \le n$-lists is the same as the number of p-subsets of a set with n elements, namely, $\binom{n}{p}$.

Proposition 9.137. *Let M be finitely generated, say, $M = \langle e_1, \ldots, e_n \rangle$. If $p \ge 1$, then $\bigwedge^p(M)$ is generated by all elements of the form e_H, where $H = i_1, \ldots, i_p$ is an increasing $p \le n$-list.*

Proof. Every element of M has some expression of the form $\sum a_i e_i$, where $a_i \in k$. We prove the proposition by induction on $p \geq 1$. Let $m_1 \wedge \cdots \wedge m_{p+1}$ be a typical generator of $\bigwedge^{p+1}(M)$. By induction,

$$m_1 \wedge \cdots \wedge m_p = \sum_H a_H e_H,$$

where $a_H \in k$ and H is an increasing $p \leq n$-list. If $m_{p+1} = \sum b_j e_j$, then

$$m_1 \wedge \cdots \wedge m_{p+1} = (\sum_H a_H e_H) \wedge (\sum_j b_j e_j).$$

Each e_j in $\sum b_j e_j$ can be moved to any position in $e_H = e_{i_1} \wedge \cdots \wedge e_{i_p}$ (with a possible change in sign) by (anti)commuting it from right to left. Of course, if $e_j = e_{i_\ell}$ for any ℓ, then this term is 0, and so we can assume that all the factors in surviving wedges are distinct and are arranged with indices in ascending order. •

Corollary 9.138. *If M can be generated by n elements, then $\bigwedge^p(M) = \{0\}$ for all $p > n$.*

Proof. Any wedge of p factors must be 0, for it must contain a repetition of one of the generators. •

Definition. If V is a free k-module of rank n, then a ***Grassmann algebra*** on V is a k-algebra $G(V)$ with identity element, denoted by e_0, such that

(a) $G(V)$ contains $\langle e_0 \rangle \oplus V$ as a submodule, where $\langle e_0 \rangle \cong k$;

(b) $G(V)$ is generated, as a k-algebra, by $\langle e_0 \rangle \oplus V$;

(c) $v^2 = 0$ for all $v \in V$;

(d) $G(V)$ is a free k-module of rank 2^n.

The computation on page 741 shows that the condition $v^2 = 0$ for all $v \in V$ implies $vu = -uv$ for all $u, v \in V$. A candidate for $G(V)$ is $\bigwedge(V)$ but, at this stage, it is not clear how to show that $\bigwedge(V)$ is free and of the desired rank.

Grassmann algebras carry a generalization of complex conjugation, and this fact is the key to proving their existence. If A is a k-algebra, then an ***algebra automorphism*** is a k-algebra isomorphism of A with itself.

Theorem 9.139. *Let V be a free k-module with basis $e_1, \ldots, e_n$, where $n \geq 1$.*

(i) *There exists a Grassmann algebra $G(V)$ with an algebra automorphism $u \mapsto \overline{u}$, called **conjugation**, such that*

$$\overline{\overline{u}} = u;$$
$$\overline{e_0} = e_0;$$
$$\overline{v} = -v \quad \textit{for all } v \in V.$$

(ii) *The Grassmann algebra $G(V)$ is a graded k-algebra*

$$G(V) = \sum_p G^p(V),$$

where

$$G^p(V) = \langle e_H : \text{ where } H \text{ is an increasing } p\text{-list}\rangle$$

[*we have extended the notation $e_H = e_{i_1} \wedge \cdots \wedge e_{i_p}$ in $\bigwedge^p(V)$ to $e_H = e_{i_1} \cdots e_{i_p}$ in $G^p(V)$*]. *Moreover, $G^p(V)$ is a free k-module with*

$$\operatorname{rank}(G^p(V)) = \binom{n}{p}.$$

Proof. (i) The proof is by induction on $n \geq 1$. The base step is clear: If $V = \langle e_1\rangle \cong k$, set $G(V) = \langle e_0\rangle \oplus \langle e_1\rangle$; note that $G(V)$ is a free k-module of rank 2. Define a multiplication on $G(V)$ by

$$e_0e_0 = e_0; \quad e_0e_1 = e_1 = e_1e_0; \quad e_1e_1 = 0.$$

It is routine to check that $G(V)$ is a k-algebra that satisfies the axioms of a Grassmann algebra. There is no choice in defining the automorphism; we must have

$$\overline{ae_0 + be_1} = ae_0 - be_1.$$

Finally, it is easy to see that $u \mapsto \overline{u}$ is the automorphism we seek.

For the inductive step, let V be a free k-module of rank $n+1$ and let $e_1, \ldots, e_{n+1}$ be a basis of V. If $W = \langle e_1, \ldots, e_n\rangle$, then the inductive hypothesis provides a Grassmann algebra $G(W)$, free of rank 2^n, and an automorphism $u \mapsto \overline{u}$ for all $u \in G(W)$. Define $G(V) = G(W) \oplus G(W)$, so that $G(V)$ is a free module of rank $2^n + 2^n = 2^{n+1}$. We make $G(V)$ into a k-algebra by defining

$$(x_1, x_2)(y_1, y_2) = (x_1y_1, x_2\overline{y}_1 + x_1y_2).$$

We now verify the four parts in the definition of Grassmann algebra.

(a) At the moment, V is not a submodule of $G(V)$. Each $v \in V$ has a unique expression of the form $v = w + ae_{n+1}$, where $w \in W$ and $a \in k$. The k-map $V \to G(V)$, given by

$$v = w + ae_{n+1} \mapsto (w, ae_0),$$

is an isomorphism of k-modules, and we identify V with its image in $G(V)$. In particular, e_{n+1} is identified with $(0, e_0)$. Note that the identity element $e_0 \in G(W)$ in $G(W)$ has been identified with $(e_0, 0)$ in $G(V)$, and that the definition of multiplication in $G(V)$ shows that $(e_0, 0)$ is the identity in $G(V)$.

(b) By induction, we know that the elements of $\langle e_0\rangle \oplus W$ generate $G(W)$ as a k-algebra; that is, all $(x_1, 0) \in G(W)$ arise from elements of W. Next, by our identification, $e_{n+1} = (0, e_0)$,

$$(x_1, 0)e_{n+1} = (x_1, 0)(0, e_0) = (0, x_1),$$

and so the elements of V generate all pairs of the form $(0, x_2)$. Since addition is coordinatewise, all $(x_1, x_2) = (x_1, 0) + (0, x_2)$ arise from V using algebra operations.

(c) If $v \in V$, then $v = w + ae_{n+1}$, where $w \in W$, and v is identified with (w, ae_0) in $G(V)$. Hence,

$$v^2 = (w, ae_0)(w, ae_0) = (w^2, ae_0\overline{w} + ae_0 w).$$

Now $w^2 = 0$, and $\overline{w} = -w$, so that $v^2 = 0$.

(d) rank $G(V) = 2^{n+1}$ because $G(V) = G(W) \oplus G(W)$.

We have shown that $G(V)$ is a Grassmann algebra. Finally, define conjugation by

$$\overline{(x_1, x_2)} = (\overline{x}_1, -\overline{x}_2).$$

The reader may check that this defines a function with the desired properties.

(ii) We prove, by induction on $n \geq 1$, that $G^p(V) = \langle e_H :$ where H is an increasing p-list$\rangle$ is a free k-module with the displayed products as a basis. The base step is obvious: If rank$(V) = 1$, say, with basis e_1, then $G(V) = \langle e_0, e_1\rangle$; moreover, both $G^0(V)$ and $G^1(V)$ are free of rank 1.

For the inductive step, assume that V is free with basis $e_1, \ldots, e_{n+1}$. As in the proof of part (i), let $W = \langle e_1 \ldots, e_n\rangle$. By induction, $G^p(W)$ is a free k-module of rank $\binom{n}{p}$ with basis all e_H, where H is an increasing $p \leq n$-list. Here are two types of elements of $G^p(V)$: elements $e_H \in G(W)$, where H is an increasing $p \leq n$-list; elements $e_H = e_{i_1} \cdots e_{i_{p-1}} e_{n+1}$, where H is an increasing $p \leq (n+1)$-list that involves e_{n+1}. We know that the elements of the first type comprise a basis of $G(W)$. The definition of multiplication in $G(V)$ gives $e_H e_{n+1} = (e_H, 0)(0, e_0) = (0, e_H)$. Thus, the number of such products is $\binom{n}{p-1}$. As $G(V) = G(W) \oplus G(W)$, we see that the union of these two types of products form a basis for $G^p(V)$, and so rank$(G^p(V)) = \binom{n}{p} + \binom{n}{p-1} = \binom{n+1}{p}$.

It remains to prove that $G^p(V)G^q(V) \subseteq G^{p+q}(V)$. Consider $e_{i_1} \cdots e_{i_p} e_{j_1} \cdots e_{j_q}$late. If some subscript i_r is the same as a subscript j_s, then this product is 0 because it has a repeated factor; if all the subscripts are distinct, then this product lies in $G^{p+q}(V)$, as desired. Therefore, $G(V)$ is a graded k-algebra whose graded part of degree p is a free k-module of rank $\binom{n}{p}$. •

Theorem 9.140 (Binomial Theorem). *If V is a free k-module of rank n, then there is an isomorphism of graded k-algebras,*

$$\bigwedge(V) \cong G(V).$$

Thus, $\bigwedge^p(V)$ is a free k-module, for all $p \geq 1$, with basis all increasing $p \leq n$-lists, hence

$$\text{rank}\Big(\bigwedge\nolimits^p(V)\Big) = \binom{n}{p}.$$

Proof. For any $p \geq 2$, consider the diagram

$$\begin{array}{ccc} \times^p V & \xrightarrow{\quad h \quad} & \bigwedge^p(V), \\ & {\scriptstyle g_p} \searrow \quad \swarrow {\scriptstyle \widehat{g}_p} & \\ & G^p(V) & \end{array}$$

where $g_p(v_1, \ldots, v_p) = v_1 \cdots v_p$. Since $v^2 = 0$ in $G^p(V)$ for all $v \in V$, the function g_p is alternating multilinear. By the universal property of exterior power, there is a (unique) k-homomorphism $\widehat{g}_p \colon \bigwedge^p(V) \to G^p(V)$ making the diagram commute; that is,

$$\widehat{g}_p(v_1 \wedge \cdots \wedge v_p) = v_1 \cdots v_p.$$

If $e_1, \ldots, e_n$ is a basis of V, then we have just seen that $G^p(V)$ is a free k-module with basis all $e_{i_1} \cdots e_{i_p}$, and so $\widehat{g}_p$ is surjective. But $\bigwedge^p(V)$ is generated by all $e_{i_1} \wedge \cdots \wedge e_{i_p}$, by Proposition 9.137. If some k-linear combination $\sum_H a_H e_H$ lies in $\ker \widehat{g}_p$, then $\sum a_H \widehat{g}_p(e_H) = 0$ in $G^p(V)$. But the list of images $\widehat{g}_p(e_H)$ forms a basis of the free k-module $G^p(V)$, so that all the coefficients $a_H = 0$. Therefore, $\ker \widehat{g}_p = \{0\}$, and so $\widehat{g}_p$ is a k-isomorphism.

Define $\gamma \colon \bigwedge(V) \to G(V)$ by $\gamma(\sum_{p=0}^n u_p) = \sum_{p=0}^n \widehat{g}_p(u_p)$, so that $\gamma(\bigwedge^p(V)) \subseteq G^p(V)$. We are done if we can show that γ is an algebra map: $\gamma(u \wedge v) = \gamma(u)\gamma(v)$. But this is clear for homogeneous elements of $\bigwedge(V)$, and hence it is true for all elements. •

Corollary 9.141. *If V is a free k-module with basis $e_1, \ldots, e_n$, then*

$$\bigwedge^n(V) = \langle e_1 \wedge \cdots \wedge e_n \rangle \cong k.$$

Proof. By Proposition 9.137, we know that $\bigwedge^n(V)$ is a cyclic module generated by $e_1 \wedge \cdots \wedge e_n$, but we cannot conclude from this proposition whether or not this element is zero. However, the binomial theorem not only says that this element is nonzero; it also says that it generates a cyclic module isomorphic to k. •

Proposition 7.43 says that if $T \colon {}_k\mathbf{Mod} \to {}_k\mathbf{Mod}$ is an additive functor, then $T(V \oplus V') \cong T(V) \oplus T(V')$. It follows, for $p \geq 2$, that $\bigwedge^p$ is *not* an additive functor: if V is a free k-module of rank n, then $\bigwedge^p(V \oplus V)$ is free of rank $\binom{2n}{p}$, whereas $\bigwedge^p(V) \oplus \bigwedge^p(V)$ is free of rank $2\binom{n}{p}$.

An astute reader will have noticed that our construction of a Grassmann algebra $G(V)$ depends not only on the free k-module V but also on a choice of basis of V. Had we chosen a second basis of V, would the second Grassmann algebra be isomorphic to the first one?

Corollary 9.142. *Let V be a free k-module, and let B and B' be bases of V. If $G(V)$ is the Grassmann algebra defined using B and if $G'(V)$ is the Grassmann algebra defined using B', then $G(V) \cong G'(V)$ as graded k-algebras.*

Proof. Both $G(V)$ and $G'(V)$ are isomorphic to $\bigwedge(V)$, and the latter has been defined without any choice of basis. •

A second proof of the binomial theorem follows from the next result.

Theorem 9.143. *For all $p \geq 0$ and all k-modules A and B, where k is a commutative ring,*

$$\bigwedge^p(A \oplus B) \cong \sum_{i=0}^{p} \Big(\bigwedge^i(A) \otimes_k \bigwedge^{p-i}(B)\Big).$$

Sketch of Proof. Let $\mathcal{A}$ be the category of all alternating anticommutative graded k-algebras $R = \sum_{p \geq 0} R^p$ ($r^2 = 0$ for all $r \in R$ of odd degree and $rs = (-1)^{pq} sr$ if $r \in R^p$ and $s \in S^q$); by Theorem 9.136, the exterior algebra $\bigwedge(A) \in \text{obj}(\mathcal{A})$ for every k-module A. If $R, S \in \text{obj}(\mathcal{A})$, then one verifies that $R \otimes_k S = \sum_{p \geq 0} \big(\sum_{i=0}^{p} R^i \otimes_k S^{p-i}\big) \in \text{obj}(\mathcal{A})$; using anticommutativity, a modest generalization of Proposition 9.101 shows that $\mathcal{A}$ has coproducts.

We claim that $(\bigwedge, D)$ is an adjoint pair of functors, where $\bigwedge \colon {}_k\textbf{Mod} \to \mathcal{A}$ sends $A \mapsto \bigwedge(A)$, and $D \colon \mathcal{A} \to {}_k\textbf{Mod}$ sends $\sum_{p \geq 0} R^p \mapsto R^1$, the terms of degree 1. If $R = \sum_p R^p$, then there is a map $\pi_R \colon \bigwedge(R^1) \to R$; define $\tau_{A,R} \colon \text{Hom}_{\mathcal{A}}(\bigwedge(A), R) \to \text{Hom}_k(A, R^1)$ by $\varphi \mapsto \pi_R(\varphi|A)$. It follows from Theorem 7.105 that $\bigwedge$ preserves coproducts; that is, $\bigwedge(A \oplus B) \cong \bigwedge(A) \otimes_k \bigwedge(B)$, and so $\bigwedge^p(A \oplus B) \cong \sum_{i=0}^{p} \Big(\bigwedge^i(A) \otimes_k \bigwedge^{p-i}(B)\Big)$. •

Here is an explicit formula for an isomorphism. In $\bigwedge^3(A \oplus B)$, we have

$$\begin{aligned}(a_1 + b_1) \wedge (a_2 + b_2) \wedge (a_3 + b_3) = {} & a_1 \wedge a_2 \wedge a_3 + a_1 \wedge b_2 \wedge a_3 \\ & + b_1 \wedge a_2 \wedge a_3 + b_1 \wedge b_2 \wedge a_3 + a_1 \wedge a_2 \wedge b_3 \\ & + a_1 \wedge b_2 \wedge b_3 + b_1 \wedge a_2 \wedge b_3 + b_1 \wedge b_2 \wedge b_3.\end{aligned}$$

By anticommutativity, this can be rewritten so that each a precedes all the b's:

$$\begin{aligned}(a_1 + b_1) \wedge (a_2 + b_2) \wedge (a_3 + b_3) = {} & a_1 \wedge a_2 \wedge a_3 - a_1 \wedge a_3 \wedge b_2 \\ & + a_2 \wedge a_3 \wedge b_1 + a_3 \wedge b_1 \wedge b_2 + a_1 \wedge a_2 \wedge b_3 \\ & + a_1 \wedge b_2 \wedge b_3 - a_2 \wedge b_1 \wedge b_3 + b_1 \wedge b_2 \wedge b_3.\end{aligned}$$

An ***i-shuffle*** is a partition of $\{1, 2, \ldots, p\}$ into two disjoint subsets $\mu_1 < \ldots < \mu_i$ and $\nu_1 < \ldots < \nu_{p-i}$; it gives the permutation $\sigma \in S_p$ with $\sigma(j) = \mu_j$ for $j \leq i$ and $\sigma(i + \ell) = \nu_\ell$ for $j = i + \ell > i$. Each "mixed" term in $(a_1 + b_1) \wedge (a_2 + b_2) \wedge (a_3 + b_3)$ gives a shuffle, with the a's giving the μ and the b's giving the ν; for example, $a_1 \wedge b_2 \wedge a_3$ is a 2-shuffle and $b_1 \wedge a_2 \wedge b_3$ is a 1-shuffle. Now $\text{sgn}(\sigma)$ counts the total number of leftward moves of a's so that they precede all the b's, and the reader may check that the signs in the rewritten expansion are $\text{sgn}(\sigma)$. Define $f \colon \bigwedge^p(A \oplus B) \to \sum_{i=0}^{p} \Big(\bigwedge^i(A) \otimes_k \bigwedge^{p-i}(B)\Big)$ by

$$f(a_1 + b_1, \ldots, a_p + b_p) = \sum_{i=0}^{p} \left(\sum_{i\text{-shuffles } \sigma} \text{sgn}(\sigma) a_{\mu_1} \wedge \cdots \wedge a_{\mu_i} \otimes b_{\nu_1} \wedge \cdots \wedge b_{\nu_{p-i}} \right).$$

Corollary 9.144. *If k is a commutative ring and V is a free k-module of rank n, then $\bigwedge^p(V)$ is a free k-module of rank $\binom{n}{p}$.*

Proof. Write $V = k \oplus B$ and use induction on $\operatorname{rank}(V)$. •

We will use exterior algebra in the next section to prove theorems about determinants, but let us first note a nice result when k is a field and, hence, k-modules are vector spaces.

Proposition 9.145. *Let k be a field, let V be a vector space over k, and let $v_1,\ldots,v_p$ be vectors in V. Then $v_1 \wedge \cdots \wedge v_p = 0$ in $\bigwedge(V)$ if and only if $v_1, \ldots, v_p$ is a linearly dependent list.*

Proof. Since k is a field, a linearly independent list $v_1, \ldots, v_p$ can be extended to a basis $v_1, \ldots, v_p, \ldots, v_n$ of V. By Corollary 9.141, $v_1 \wedge \cdots \wedge v_n \neq 0$. But $v_1 \wedge \cdots \wedge v_p$ is a factor of $v_1 \wedge \cdots \wedge v_n$, so that $v_1 \wedge \cdots \wedge v_p \neq 0$.

Conversely, if $v_1, \ldots, v_p$ is linearly dependent, there is some i with $v_i = \sum_{j \neq i} a_j v_j$, where $a_j \in k$. Hence,

$$
\begin{aligned}
v_1 \wedge \cdots \wedge v_i \wedge \cdots \wedge v_p &= v_1 \wedge \cdots \wedge \sum_{j \neq i} a_j v_j \wedge \cdots \wedge v_p \\
&= \sum_{j \neq i} a_j v_1 \wedge \cdots \wedge v_j \wedge \cdots \wedge v_p.
\end{aligned}
$$

After expanding, each term has a repeated factor v_j, and so this is 0. •

We introduced exterior algebra, at the beginning of this section, by looking at Jacobians; we now end this section by applying exterior algebra to differential forms. Let X be an open connected[21] subset of euclidean space $\mathbb{R}^n$. A function $f : X \to \mathbb{R}$ is called a ***C^∞-function*** if, for all $p \geq 1$, the pth partials $\partial^p f / \partial^p x_i$ exist for all $i = 1, \ldots, n$, as do all the mixed partials.

Definition. If X is a connected open subset of $\mathbb{R}^n$, define

$$A(X) = \{ f : X \to \mathbb{R} : f \text{ is a } C^\infty\text{-function} \}.$$

The condition that X be a connected open subset of $\mathbb{R}^n$ is present so that C^∞-functions are defined. It is easy to see that $A(X)$ is a commutative ring under pointwise operations:

$$f + g : x \mapsto f(x) + g(x); \qquad fg : x \mapsto f(x)g(x).$$

In the free $A(X)$-module $A(X)^n$ of all n-tuples, rename the standard basis

$$dx_1, \ldots, dx_n.$$

[21]A subset X is ***open*** if, for each $x \in X$, there is some $r > 0$ so that all points y with distance $|y - x| < r$ also lie in X. An open subset X of $\mathbb{R}^n$ is ***connected*** if we can join any pair of points in X by a path lying wholly in X.

By the binomial theorem, each element $\omega \in \bigwedge^p(A(X)^n)$ has a unique expression

$$\omega = \sum_{i_1 \ldots i_p} f_{i_1 \ldots i_p} dx_{i_1} \wedge \cdots \wedge dx_{i_p},$$

where $f_{i_1 \ldots i_p} \in A(X)$ is a C^∞-function on X and $i_1 \ldots i_p$ is an increasing $p \le n$-list. We write

$$\Omega^p(X) = \bigwedge^p(A(X)^n),$$

and we call its elements ***differential*** p***-forms*** on X.

Definition. The ***exterior derivative*** $d^p \colon \Omega^p(X) \to \Omega^{p+1}(X)$ is defined as follows:

(i) If $f \in \Omega^0(X) = A(X)$, then $d^0 f = \sum_{j=1}^n \frac{\partial f}{\partial x_j} dx_j$;

(ii) If $p \ge 1$ and $\omega \in \Omega^p(X)$, then $\omega = \sum_{i_1 \ldots i_p} f_{i_1 \ldots i_p} dx_{i_1} \wedge \cdots \wedge dx_{i_p}$, and we define

$$d^p \omega = \sum_{i_1 \ldots i_p} d^0(f_{i_1 \ldots i_p}) \wedge dx_{i_1} \wedge \cdots \wedge dx_{i_p}.$$

If X is an open connected subset of $\mathbb{R}^n$, the exterior derivatives give a sequence of $A(X)$-maps, called the ***de Rham complex***:

$$0 \to \Omega^0(X) \xrightarrow{d^0} \Omega^1(X) \xrightarrow{d^1} \cdots \xrightarrow{d^{n-1}} \Omega^n(X) \to 0.$$

Proposition 9.146. *If X is a connected open subset of $\mathbb{R}^n$, then*

$$d^{p+1} d^p \colon \Omega^p(X) \to \Omega^{p+2}(X)$$

is the zero map for all $p \ge 0$.

Proof. It suffices to prove that $dd\omega = 0$, where $\omega = f dx_I$ (we are using an earlier abbreviation: $dx_I = dx_{i_1} \wedge \cdots \wedge dx_{i_p}$, where $I = i_1, \ldots, i_p$ is an increasing $p \le n$-list).

Now

$$\begin{aligned} dd\omega &= d(d^0 f \wedge x_I) \\ &= d\left(\sum_i \frac{\partial f}{\partial x_i} dx_i \wedge dx_I\right) \\ &= \sum_i \sum_j \frac{\partial^2 f}{\partial x_i \partial x_j} dx_j \wedge dx_i \wedge dx_I. \end{aligned}$$

Compare the i, j and j, i terms in this double sum: The first is

$$\frac{\partial^2 f}{\partial x_i \partial x_j} dx_j \wedge dx_i \wedge dx_I;$$

the second is

$$\frac{\partial^2 f}{\partial x_j \partial x_i}\, dx_i \wedge dx_j \wedge dx_I.$$

But these cancel, for the mixed second partials are equal:

$$dx_i \wedge dx_j = -dx_j \wedge dx_i. \quad \bullet$$

Example 9.147.
Consider the special case of the de Rham complex for $n = 3$.

$$0 \to \Omega^0(X) \xrightarrow{d^0} \Omega^1(X) \xrightarrow{d^1} \Omega^2(X) \xrightarrow{d^2} \Omega^3(X) \to 0$$

If $\omega \in \Omega^0(X)$, then $\omega = f(x, y, z) \in A(X)$, and

$$d^0 f = \frac{\partial f}{\partial x}\, dx + \frac{\partial f}{\partial y}\, dy + \frac{\partial f}{\partial z}\, dz,$$

a 1-form resembling grad(f).

If $\omega \in \Omega^1(X)$, then $\omega = f dx + g dy + h dz$, and a simple calculation gives $d^1\omega =$

$$\left(\frac{\partial g}{\partial x} - \frac{\partial f}{\partial y}\right) dx \wedge dy + \left(\frac{\partial h}{\partial y} - \frac{\partial g}{\partial z}\right) dy \wedge dz + \left(\frac{\partial f}{\partial z} - \frac{\partial h}{\partial x}\right) dz \wedge dx,$$

a 2-form resembling curl(ω).

If $\omega \in \Omega^2(X)$, then $\omega = F dy \wedge dz + G dz \wedge dx + H dx \wedge dy$. Now

$$d^2\omega = \frac{\partial F}{\partial x} + \frac{\partial G}{\partial y} + \frac{\partial H}{\partial z},$$

a 3-form resembling div(ω).

These are not mere resemblances. Since $\Omega^1(X)$ is a free $A(X)$-module with basis dx, dy, dz, we see that $d^0\omega$ is grad(ω) when ω is a 0-form. Now $\Omega^2(X)$ is a free $A(X)$-module, but we now choose a basis

$$dx \wedge dy, dy \wedge dz, dz \wedge dx$$

instead of the usual basis $dx \wedge dy, dx \wedge dz, dy \wedge dz$; it follows that $d^1\omega$ is curl(ω) in this case. Finally, $\Omega^3(X)$ has a basis $dx \wedge dy \wedge dz$, and so $d^3\omega$ is div(ω) when ω is a 2-form. We have shown that the de Rham complex is

$$0 \to \Omega^0(X) \xrightarrow{\text{grad}} \Omega^1(X) \xrightarrow{\text{curl}} \Omega^2(X) \xrightarrow{\text{div}} \Omega^3(X) \to 0.$$

Proposition 9.146 now gives the familiar identities from advanced calculus:

$$\text{curl} \cdot \text{grad} = 0 \quad \text{and} \quad \text{div} \cdot \text{curl} = 0.$$

We call a 1-form ω ***closed*** if $d\omega = 0$, and we call it ***exact*** if $\omega = \operatorname{grad} f$ for some C^∞-function f. More generally, call a p-form ω ***closed*** if $d^p\omega = 0$, and call it ***exact*** if $\omega = d^{p-1}\omega'$ for some $(p-1)$-form ω'. Thus, $\omega \in \Omega^p(X)$ is closed if and only if $\omega \in \ker d^p$ and ω is exact if and only if $\omega \in \operatorname{im} d^{p-1}$. Therefore, the de Rham complex is an exact sequence of $A(X)$-modules if and only if every closed form is exact; this is the etymology of the adjective *exact* in "exact sequence." It can be proved that the de Rham complex is an exact sequence whenever X is a simply connected open subset of $\mathbb{R}^n$. For any (not necessarily simply connected) space X, we have $\operatorname{im}\operatorname{grad} \subseteq \ker \operatorname{curl}$ and $\operatorname{im}\operatorname{curl} \subseteq \ker \operatorname{div}$, and the $\mathbb{R}$-vector spaces $\ker\operatorname{curl}/\operatorname{im}\operatorname{grad}$ and $\ker\operatorname{div}/\operatorname{im}\operatorname{curl}$ are called the *cohomology groups* of X. ◄

EXERCISES

9.85 Let $G(V)$ be the Grassmann algebra of a free k-module V, and let $u = \sum_p u_p \in G(V)$, where $u_p \in G^p(V)$ is homogeneous of degree p. If $\overline{u}$ is the conjugate of u in Theorem 9.139, prove that $\overline{u} = \sum_p (-1)^p u_p$.

9.86 (i) Let p be a prime. Show that $\bigwedge^2(\mathbb{I}_p \oplus \mathbb{I}_p) \neq 0$, where $\mathbb{I}_p \oplus \mathbb{I}_p$ is viewed as a $\mathbb{Z}$-module (i.e., as an abelian group).

(ii) Let $D = \mathbb{Q}/\mathbb{Z} \oplus \mathbb{Q}/\mathbb{Z}$. Prove that $\bigwedge^2(D) = 0$, and conclude that if $i : \mathbb{I}_p \oplus \mathbb{I}_p \to D$ is an inclusion, then $\bigwedge^2(i)$ is not an injection.

9.87 (i) If k is a commutative ring and N is a direct summand of a k-module M, prove that $\bigwedge^p(N)$ is a direct summand of $\bigwedge^p(M)$ for all $p \geq 0$.

Hint. Use Corollary 7.17 on page 434.

(ii) If k is a field and $i : W \to V$ is an injection of vector spaces over k, prove that $\bigwedge^p(i)$ is an injection for all $p \geq 0$.

9.88 Prove, for all p, that the functor $\bigwedge^p$ preserves surjections.

9.89 If P is a projective k-module, where k is a commutative ring, prove that $\bigwedge^q(P)$ is a projective k-module for all q.

9.90 Let k be a field, and let V be a vector space over k. Prove that two linearly independent lists $u_1, \ldots, u_p$ and $v_1, \ldots, v_p$ span the same subspace of V if and only if there is a nonzero $c \in k$ with $u_1 \wedge \cdots \wedge u_p = cv_1 \wedge \cdots \wedge v_p$.

9.91 If U and V are R-modules over a commutative ring R and if $U' \subseteq U$ and $V' \subseteq V$ are submodules, prove that

$$(U \otimes_R V)/(U' \otimes_R V + U \otimes_R V') \cong (U/U') \otimes_R V \oplus U \otimes_R (V/V').$$

Hint. Compute the kernel and image of $\varphi\colon U \otimes_R V \to (U/U') \otimes_R V \oplus U \otimes_R (V/V')$ defined by $\varphi\colon u \otimes v \mapsto (u + U') \otimes v + u \otimes (v + V')$.

9.92 Define the ***symmetric algebra*** on a k-module M to be $S(M) = T(M)/I$, where I is the two-sided ideal generated by all $m \otimes m' - m' \otimes m$, where $m, m' \in M$.

(i) Prove that I is a graded ideal, so that $S(M)$ is a graded k-algebra.

(ii) Prove that $S(M)$ is commutative.

9.93 (i) Define a ***free commutative k-algebra***, and prove that if M is the free k-module with basis X, then $S(M)$ is the free commutative k-algebra on X. Conclude that $S(M)$ is independent of the choice of basis of the free k-module M.

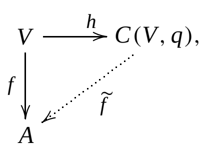
(ii) Define $k[X]$ to be the polynomial ring in commuting variables X if every $u \in k[X]$ has a unique expression as a polynomial in finitely many elements of X. Prove that if M is the free k-module with basis X, then $S(M)$ is the polynomial ring in commuting variables X.[22]

(iii) Prove that if M is a free k-module of finite rank n, then $S^p(M)$ is a free k-module of rank $\binom{n+p-1}{p}$.

Hint. Use the combinatorial fact that there are $\binom{n+p-1}{p}$ ways to distribute p identical objects among n boxes.

(iv) Prove that every commutative k-algebra is a quotient of a free commutative k-algebra.

9.94 Let V be a finite-dimensional vector space over a field k, and let $q \colon V \to k$ be a quadratic form on V. Define the ***Clifford algebra*** $C(V, q)$ as the quotient $C(V, q) = T(V)/J$, where J is the two-sided ideal generated by all elements of the form $v \otimes v - q(v)1$ (note that J is not a graded ideal). For $v \in V$, denote the coset $v + J$ by $[v]$, and define $h \colon V \to C(V, q)$ by $h(v) = [v]$. Prove that $C(V, q)$ is a solution to the following universal problem:



where A is a k-algebra and $f \colon V \to A$ is a k-module map with $f(v)^2 = q(v)$ for all $v \in V$.

If $\dim(V) = n$ and q is nondegenerate, then it can be proved that $\dim(C(V, q)) = 2^n$. In particular, if $k = \mathbb{R}$ and $n = 2$, then the Clifford algebra has dimension 4 and $C(v, q) \cong \mathbb{H}$, the division ring of quaternions. Clifford algebras are used in the study of quadratic forms, hence of orthogonal groups; see Jacobson, *Basic Algebra* II, pp. 228–245.

9.9 Determinants

We have been using familiar properties of determinants, even though the reader may have seen their verifications only over a field and not over a general commutative ring. Since determinants of matrices whose values lie in a commutative ring are of interest, the time has come to establish these properties in general, for exterior algebra is now available to help us.

If k is a commutative ring, we claim that every k-module map $\gamma \colon k \to k$ is just multiplication by some $d \in k$: If $\gamma(1) = d$, then

$$\gamma(a) = \gamma(a1) = a\gamma(1) = ad = da$$

[22] In the fourth section of Chapter 6, we assumed the existence of this big polynomial ring in order to construct the algebraic closure of a field.

Our earlier definition of $k[x, y]$ as $R[y]$, where $R = k[x]$, was careless. For example, it does not follow that $k[x, y] = k[y, x]$, although these two rings are isomorphic. However, if M is the free k-module with basis x, y, then y, x is also a basis of k-algebra M, and so $k[x, y] = k[y, x]$.

for all $a \in k$, because γ is a k-module map. Here is a slight generalization. If $V = \langle v \rangle \cong k$, then every k-map $\gamma\colon V \to V$ has the form $\gamma\colon av \mapsto dav$, where $\gamma(v) = dv$. Suppose now that V is a free k-module with basis $e_1, \ldots, e_n$; Corollary 9.141 shows that $\bigwedge^n(V)$ is free of rank 1 with generator $e_1 \wedge \cdots \wedge e_n$. It follows that every k-map $\gamma\colon \bigwedge^n(V) \to \bigwedge^n(V)$ has the form $\gamma(a(e_1 \wedge \cdots \wedge e_n)) = d(a(e_1 \wedge \cdots \wedge e_n))$.

Definition. If V is a free k-module with basis $e_1, \ldots, e_n$, and if $f\colon V \to V$ is a k-homomorphism, then the ***determinant*** of f, denoted by $\det(f)$, is the element $\det(f) \in k$ for which

$$\begin{aligned}\textstyle\bigwedge^n(f)\colon e_1 \wedge \cdots \wedge e_n \mapsto\ & f(e_1) \wedge \cdots \wedge f(e_n)\\ &= \det(f)(e_1 \wedge \cdots \wedge e_n).\end{aligned}$$

If $A = [a_{ij}]$ is an $n \times n$ matrix with entries in k, then A defines a k-map $f\colon k^n \to k^n$ by $f(x) = Ax$, where $x \in k^n$ is a column vector. If $e_1, \ldots, e_n$ is the standard basis of k^n, then $f(e_i) = \sum_j a_{ji} e_j$, and the matrix $A = [a_{ij}]$ associated to f has ith column the coordinates of $f(e_i) = Ae_i$. We define $\det(A) = \det(f)$:

$$Ae_1 \wedge \cdots \wedge Ae_n = \det(A)(e_1 \wedge \cdots \wedge e_n).$$

Thus, the wedge of the columns of A in $\bigwedge^n(k^n)$ is a constant multiple of $e_1 \wedge \cdots \wedge e_n$, and $\det(A)$ is that constant.

Example 9.148.
If

$$A = \begin{bmatrix} a & c \\ b & d \end{bmatrix},$$

then the wedge product of the columns of A is

$$\begin{aligned}(ae_1 + be_2) \wedge (ce_1 + de_2) &= ace_1 \wedge e_1 + ade_1 \wedge e_2 + bce_2 \wedge e_1 + bde_2 \wedge e_2\\ &= ade_1 \wedge e_2 + bce_2 \wedge e_1\\ &= ade_1 \wedge e_2 - bce_1 \wedge e_2\\ &= (ad - bc)(e_1 \wedge e_2).\end{aligned}$$

Therefore, $\det(A) = ad - bc$. ◂

The reader has probably noticed that this calculation is a repetition of the calculation on page 742 where we computed the Jacobian of a change of variables in a double integral. The next example considers triple integrals.

Example 9.149.
Let us change variables in a triple integral $\iiint_D f(x, y, z)dxdydz$ using equations:

$$\begin{aligned}x &= F(u, v, w);\\ y &= G(u, v, w);\\ z &= H(u, v, w).\end{aligned}$$

Denote a basis of the tangent space T of $f(x, y, z)$ at a point $P = (x_0, y_0, z_0)$ by dx, dy, dz. T If du, dv, dw is another basis of T, then the chain rule defines a linear transformation on T by the equations:

$$\begin{aligned} dx &= F_u du + F_v dv + F_w dw \\ dy &= G_u du + G_v dv + G_w dw \\ dz &= H_u du + H_v dv + H_w dw. \end{aligned}$$

If we write the differential $dxdydz$ in the integrand as $dx \wedge dy \wedge dz$, then the change of variables gives the new differential

$$dx \wedge dy \wedge dz = \det\left(\begin{bmatrix} F_u & F_v & F_w \\ G_u & G_v & G_w \\ H_u & H_v & H_w. \end{bmatrix}\right) du \wedge dv \wedge dw:$$

expand

$$(F_u du + F_v dv + F_w dw) \wedge (G_u du + G_v dv + G_w dw) \wedge (H_u du + H_v dv + H_w dw)$$

to obtain nine terms, three of which involve $(du)^2$, $(dv)^2$, or $(dw)^2$, and hence are 0. Of the remaining six terms, three have a minus sign, and it is now easy to see that this sum is the determinant. ◂

Proposition 9.150. *Let k be a commutative ring.*

(i) *If I is the identity matrix, then* $\det(I) = 1$.

(ii) *If A and B are $n \times n$ matrices with entries in k, then*

$$\det(AB) = \det(A)\det(B).$$

Proof. Both results follow from $\bigwedge^n$ being a functor on ${}_k\mathbf{Mod}$.

(i) The linear transformation corresponding to the identity matrix is 1_{k^n}, and every functor takes identities to identities.

(ii) If f and g are the linear transformations on k^n arising from A and B, respectively, then fg is the linear transformation arising from AB. If we denote $e_1 \wedge \cdots \wedge e_n$ by e_N, then

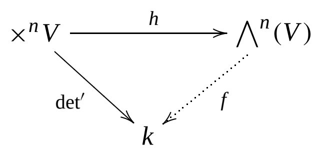
$$\begin{aligned} \det(fg)e_N &= \bigwedge^n(fg)(e_N) \\ &= \bigwedge^n(f)\Big(\bigwedge^n(g)(e_N)\Big) \\ &= \bigwedge^n(f)(\det(g)e_N) \\ &= \det(g)\bigwedge^n(f)(e_N) \\ &= \det(f)\det(g)e_N; \end{aligned}$$

the next to last equation uses the fact that $\bigwedge^n(f)$ is a k-map. Therefore,

$$\det(AB) = \det(fg) = \det(f)\det(g) = \det(A)\det(B). \quad \bullet$$

Corollary 9.151. *If k is a commutative ring, then* $\det\colon \mathrm{Mat}_n(k) \to k$ *is the unique alternating multilinear function with* $\det(I) = 1$.

Proof. The definition of determinant (as the wedge of the columns) shows that it is an alternating multilinear function $\det\colon \times^n V \to k$, where $V = k^n$, and the proposition shows that $\det(I) = 1$. The uniqueness of such a function follows from the universal property of $\bigwedge^n$.



If $\det'$ is multilinear, then there exists a k-map $f\colon \bigwedge^n(V) \to k$ with $fh = \det'$; if $\det'(e_1, \ldots, e_n) = 1$, then $f(e_1 \wedge \cdots \wedge e_n) = 1$. Since $\bigwedge^n(V) \cong k$, every k-map $f\colon \bigwedge^n(V) \to k$ is determined by $f(e_1 \wedge \cdots \wedge e_n)$. Thus, the map f is the same for $\det'$ as it is for det, and so $\det' = fh = \det$. $\bullet$

We now show that the determinant just defined coincides with the familiar determinant function.

Lemma 9.152. *Let $e_1, \ldots, e_n$ be a basis of a free k-module, where k is a commutative ring. If σ is a permutation of $1, 2, \ldots, n$, then*

$$e_{\sigma(1)} \wedge \cdots \wedge e_{\sigma(n)} = \mathrm{sgn}(\sigma)(e_1 \wedge \cdots \wedge e_n).$$

Proof. Since $m \wedge m' = -m' \wedge m$, it follows that interchanging adjacent factors in the product $e_1 \wedge \cdots \wedge e_n$ gives

$$e_1 \wedge \cdots \wedge e_i \wedge e_{i+1} \wedge \cdots \wedge e_n = -e_1 \wedge \cdots \wedge e_{i+1} \wedge e_i \wedge \cdots \wedge e_n.$$

More generally, if $i < j$, then we can interchange e_i and e_j by a sequence of interchanges of adjacent factors, each of which causes a sign change. By Exercise 2.7 on page 50, this can be accomplished with an odd number of interchanges of adjacent factors. Hence, for any transposition $\tau \in S_n$, we have

$$\begin{aligned} e_{\tau(1)} \wedge \cdots \wedge e_{\tau(n)} &= e_1 \wedge \cdots \wedge e_j \wedge \cdots \wedge e_i \wedge \cdots \wedge e_n \\ &= -[e_1 \wedge \cdots \wedge e_i \wedge \cdots \wedge e_j \wedge \cdots \wedge e_n] \\ &= \mathrm{sgn}(\tau)(e_1 \wedge \cdots \wedge e_n). \end{aligned}$$

We prove the general statement by induction on m, where σ is a product of m transpositions. The base step having just been proven, we proceed to the inductive step. Write $\sigma = \tau_1\tau_2 \cdots \tau_{m+1}$, and denote $\tau_2 \cdots \tau_{m+1}$ by σ'. By the inductive hypothesis,

$$e_{\sigma'(1)} \wedge \cdots \wedge e_{\sigma'(n)} = \mathrm{sgn}(\sigma')e_1 \wedge \cdots \wedge e_n,$$

and so

$$\begin{aligned} e_{\sigma(1)} \wedge \cdots \wedge e_{\sigma(n)} &= e_{\tau_1\sigma'(1)} \wedge \cdots \wedge e_{\tau_1\sigma'(n)} \\ &= -e_{\sigma'(1)} \wedge \cdots \wedge e_{\sigma'(n)} \qquad \text{(base step)} \\ &= -\operatorname{sgn}(\sigma')(e_1 \wedge \cdots \wedge e_n) \qquad \text{(inductive step)} \\ &= \operatorname{sgn}(\tau_1)\operatorname{sgn}(\sigma')(e_1 \wedge \cdots \wedge e_n) \\ &= \operatorname{sgn}(\sigma)(e_1 \wedge \cdots \wedge e_n). \quad \bullet \end{aligned}$$

Remark. There is a simpler proof of this lemma in the special case when k is a field. If k has characteristic 2, then Lemma 9.152 is obviously true, and so we may assume that characteristic k is not 2. Let $e_1, \ldots, e_n$ be the standard basis of k^n. If $\sigma \in S_n$, define a linear transformation $\varphi_\sigma \colon k^n \to k^n$ by $\varphi_\sigma \colon e_i \mapsto e_{\sigma(i)}$. Since $\varphi_{\sigma\tau} = \varphi_\sigma\varphi_\tau$, as is easily verified, there is a group homomorphism $d \colon S_n \to k^\times$ given by $d \colon \sigma \mapsto \det(\varphi_\sigma)$. If σ is a transposition, then $\sigma^2 = (1)$ and $d(\sigma)^2 = 1$ in $k^\times$. Since k is a field, $d(\sigma) = \pm 1$. As every permutation is a product of transpositions, it follows that $d(\sigma) = \pm 1$ for every permutation σ, and so $\operatorname{im}(d) \le \{\pm 1\}$. Now there are only two homomorphisms $S_n \to \{\pm 1\}$: the trivial homomorphism with kernel S_n and sgn. To show that $d = \operatorname{sgn}$, it suffices to show $d((1\ \ 2)) \ne 1$. But $d((1\ \ 2)) = \det(\varphi_{(1\,2)})$; that is,

$$\begin{aligned} \det(\varphi_{(1\,2)})(e_1 \wedge \cdots \wedge e_n) &= \varphi_{(1\,2)}(e_1) \wedge \cdots \wedge \varphi_{(1\,2)}(e_n) \\ &= e_2 \wedge e_1 \wedge e_3 \wedge \cdots \wedge e_n \\ &= -(e_1 \wedge \cdots \wedge e_n). \end{aligned}$$

Therefore, $d((1\ \ 2)) = -1 \ne 1$, because k does not have characteristic 2, and so, for all $\sigma \in S_n$, $d(\sigma) = \det(\varphi_\sigma) = \operatorname{sgn}(\sigma)$; that is, $e_{\sigma(1)} \wedge \cdots \wedge e_{\sigma(n)} = \operatorname{sgn}(\sigma)(e_1 \wedge \cdots \wedge e_n)$. ◂

Proposition 9.153 (Complete Expansion). *Let $e_1, \ldots, e_n$ be a basis of a free k-module, where k is a commutative ring. If $A = [a_{ij}]$ is an $n \times n$ matrix with entries in k, then*

$$\det(A) = \sum_{\sigma \in S_n} \operatorname{sgn}(\sigma) a_{\sigma(1),1} a_{\sigma(2),2} \cdots a_{\sigma(n),n}.$$

Proof. Expand the wedge of the columns of A:

$$\begin{aligned} &\sum_{j_1} a_{j_1 1} e_{j_1} \wedge \sum_{j_2} a_{j_2 2} e_{j_2} \wedge \cdots \wedge \sum_{j_n} a_{j_n n} e_{j_n} \\ &\qquad\qquad = \sum_{j_1, j_2, \ldots, j_n} a_{j_1 1} e_{j_1} \wedge a_{j_2 2} e_{j_2} \wedge \cdots \wedge a_{j_n n} e_{j_n}. \end{aligned}$$

Any summand in which $e_{j_p} = e_{j_q}$ must be 0, for it has a repeated factor, and so we may assume, in any surviving term, that $j_1, j_2, \ldots, j_n$ are all distinct; that is, there is some

permutation $\sigma \in S_n$ with $j_r = \sigma(r)$ when $1 \le r \le n$. The original product now has the form

$$\sum_{\sigma \in S_n} a_{\sigma(1)1} a_{\sigma(2)2} \cdots a_{\sigma(n)n} e_{\sigma(1)} \wedge e_{\sigma(2)} \wedge \cdots \wedge e_{\sigma(n)}.$$

By the lemma, $e_{\sigma(1)} \wedge e_{\sigma(2)} \wedge \cdots \wedge e_{\sigma(n)} = \operatorname{sgn}(\sigma)(e_1 \wedge \cdots \wedge e_n)$. Therefore, the wedge of the columns is equal to $\Big(\sum_{\sigma \in S_n} \operatorname{sgn}(\sigma) a_{\sigma(1)1} a_{\sigma(2)2} \cdots a_{\sigma(n)n}\Big)(e_1 \wedge \cdots \wedge e_n)$, and this completes the proof. •

Quite often, the complete expansion is taken as the definition of the determinant.

Corollary 9.154. *Let k be a commutative ring, and let A be an $n \times n$ matrix with entries in k. If $u \in k$, then* $\det(uI - A) = f(u)$*, where $f(x) \in k[x]$ is a monic polynomial of degree n. Moreover, the coefficient of x^{n-1} in $f(x)$ is* $-\operatorname{tr}(A)$.

Proof. Let $A = [a_{ij}]$ and let $B = [b_{ij}]$, where $b_{ij} = u\delta_{ij} - a_{ij}$ (where δ_{ij} is the Kronecker delta). By the proposition,

$$\det(B) = \sum_{\sigma \in S_n} \operatorname{sgn}(\sigma) b_{\sigma(1),1} b_{\sigma(2),2} \cdots b_{\sigma(n),n}.$$

If $\sigma = (1)$, then the corresponding term in the complete expansion is

$$b_{11} b_{22} \cdots b_{nn} = \prod_i (u - a_{ii}) = g(u),$$

where $g(x) = \prod_i (x - a_{ii})$ is a monic polynomial in $k[x]$ of degree n. If $\sigma \ne (1)$, then the σth term in the complete expansion cannot have exactly $n - 1$ factors from the diagonal of $uI - A$, for if σ fixes $n - 1$ indices, then $\sigma = (1)$. Therefore, the sum of the terms over all $\sigma \ne (1)$ is either 0 or a polynomial in $k[x]$ of degree at most $n - 2$. It follows that $\deg(f) = n$ and that the coefficient of x^{n-1} is $-\sum_i a_{ii} = -\operatorname{tr}(A)$. •

Proposition 9.155. *If A is an $n \times n$ matrix with entries in a commutative ring k, then*

$$\det(A^t) = \det(A),$$

where A^t is the transpose of A.

Proof. If $A = [a_{ij}]$, write the complete expansion of $\det(A)$ more compactly:

$$\det(A) = \sum_{\sigma \in S_n} \operatorname{sgn}(\sigma) \prod_i a_{\sigma(i),i}.$$

For any permutation $\tau \in S_n$, we have $i = \tau(j)$ for all i, and so

$$\prod_i a_{\sigma(i),i} = \prod_j a_{\sigma(\tau(j)),\tau(j)},$$

for this merely rearranges the factors in the product. Choosing $\tau = \sigma^{-1}$ gives

$$\prod_j a_{\sigma(\tau(j)),\tau(j)} = \prod_j a_{j,\sigma^{-1}(j)}.$$

Therefore,

$$\det(A) = \sum_{\sigma \in S_n} \operatorname{sgn}(\sigma) \prod_j a_{j,\sigma^{-1}(j)}.$$

Now $\operatorname{sgn}(\sigma) = \operatorname{sgn}(\sigma^{-1})$ [if $\sigma = \tau_1 \cdots \tau_q$, where the τ are transpositions, then $\sigma^{-1} = \tau_q \cdots \tau_1$]; moreover, as σ varies over S_n, so does σ^{-1}. Hence, writing $\sigma^{-1} = \rho$ gives

$$\det(A) = \sum_{\rho \in S_n} \operatorname{sgn}(\rho) \prod_j a_{j,\rho(j)}.$$

Now write $A^t = [b_{ij}]$, where $b_{ij} = a_{ji}$. Then

$$\det(A^t) = \sum_{\rho \in S_n} \operatorname{sgn}(\rho) \prod_j b_{\rho(j),j} = \sum_{\rho \in S_n} \operatorname{sgn}(\rho) \prod_j a_{j,\rho(j)} = \det(A). \quad \bullet$$

We have already seen that the eigenvalues $\alpha_1, \ldots, \alpha_n$ of an $n \times n$ matrix A, with entries in a field k, are the roots of the characteristic polynomial

$$\psi_A(x) = \det(xI - A) \in k[x].$$

We have seen that $\det(A) = \prod_i \alpha_i$, and we are now going to see that $\operatorname{tr}(A) = \sum_i \alpha_i$.

Proposition 9.156. *If $A = [a_{ij}]$ is an $n \times n$ matrix with entries in a field k, then*

$$\operatorname{tr}(A) = \alpha_1 + \alpha_2 + \cdots + \alpha_n.$$

Proof. In the complete expansion of $\det(xI - A)$, the diagonal corresponds to the term $(x - a_{\sigma(1),1})(x - a_{\sigma(2),2}) \cdots (x - a_{\sigma(n),n})$ with σ the identity permutation. If $\sigma \neq 1$, then there are at least two terms off the diagonal, and so the degree of this term is at most $n-2$. Therefore, the coefficient b_{n-1} of x^{n-1} in the diagonal term, namely, $-\sum_i a_{ii}$, coincides with the coefficient of x^{n-1} in $\psi_A(x)$, namely, $-\sum_i \alpha_i = -\operatorname{tr}(A)$, where $\alpha_1, \ldots, \alpha_n$ are the eigenvalues of A. On the other hand,

$$\psi_A(x) = \prod_i (x - \alpha_i),$$

and so the coefficient of x^{n-1} is $-\sum_i \alpha_i$, as desired. $\bullet$

We know that similar matrices have the same determinant and the same trace. The next corollary generalizes these facts, for all the coefficients of their characteristic polynomials coincide.

Corollary 9.157. *If A and B are similar $n \times n$ matrices with entries in a field k, then A and B have the same characteristic polynomial.*

Proof. There is a nonsingular matrix P with $B = PAP^{-1}$, and

$$\begin{aligned} \psi_B(x) &= \det(xI - B) \\ &= \det(xI - PAP^{-1}) \\ &= \det\Big(P(xI - A)P^{-1}\Big) \\ &= \det(P)\det(xI - A)\det(P)^{-1} \\ &= \det(xI - A) \\ &= \psi_A(x). \quad \bullet \end{aligned}$$

Definition. Let A be an $n \times n$ matrix with entries in a commutative ring k. If $H = i_1, \ldots, i_p$ and $L = j_1, \ldots, j_p$ are increasing $p \le n$-lists, then A_{HL} is the $p \times p$ ***submatrix*** $[a_{st}]$, where $(s, t) \in H \times L$. A ***minor of order*** p is the determinant of a $p \times p$ submatrix.

For example, every entry a_{ij} is a minor of $A = [a_{ij}]$ of order 1. If

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix},$$

then some minors of order 2 are

$$\det \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \text{ and } \det \begin{bmatrix} a_{12} & a_{13} \\ a_{32} & a_{33} \end{bmatrix}.$$

In particular, if $1 \le i \le n$, let i' denote the increasing $n - 1 \le n$-list in which i is omitted; thus, an $(n-1) \times (n-1)$ submatrix has the form $A_{i'j'}$, and its determinant is a minor of order $n - 1$. Note that $A_{i'j'}$ is the submatrix obtained from A by deleting its ith row and jth column.

Lemma 9.158. *Let k be a commutative ring, and let $x_{i_1}, \ldots, x_{i_p} \in k^n$ be regarded as columns of an $n \times p$ matrix A, where $H = i_1, \ldots, i_p$ is an increasing $p \le n$-list. Then*

$$x_{i_1} \wedge \cdots \wedge x_{i_p} = \sum_L \det(A_{L,H}) e_L,$$

where L varies over all increasing $p \le n$-lists.

Proof. For $\ell = 1, 2, \ldots, p$, write $x_{i_\ell} = \sum_{t_\ell} a_{t_\ell i_\ell} e_{t_\ell}$, so that

$$\begin{aligned} x_{i_1} \wedge \cdots \wedge x_{i_p} &= \sum_{t_1} a_{t_1 i_1} e_{t_1} \wedge \cdots \wedge \sum_{t_p} a_{t_p i_p} e_{t_p} \\ &= \sum_{t_1 \ldots t_p} a_{t_1 i_1} \cdots a_{t_p i_p} e_{t_1} \wedge \cdots \wedge e_{t_p}. \end{aligned}$$

All terms involving a repeated index are 0, so that we may assume that the sum is over all $t_1 \dots t_p$ having no repetitions; that is, over all p-sublists $T = \{t_1 \dots t_p\}$ of $\{1, 2, \dots, n\}$. Collecting terms, we may rewrite the sum as

$$\sum_T \sum_{T=\{t_1,\dots,t_p\}} a_{t_1 i_1} \cdots a_{t_p i_p} e_{t_1} \wedge \cdots \wedge e_{t_p}.$$

For any fixed p-sublist $T = \{t_1, \dots, t_p\}$, let $L = \ell_1, \ell_2, \dots, \ell_p$ be the increasing p-list consisting of the integers in T; thus, there is a permutation $\sigma \in S_p$ with $\ell_{\sigma(1)} = t_1, \dots, \ell_{\sigma(p)} = t_p$. With this notation,

$$\begin{aligned}\sum_{T=\{t_1,\dots,t_p\}} a_{t_1 i_1} \cdots a_{t_p i_p} (e_{t_1} \wedge \cdots \wedge e_{t_p}) &= \sum_{\sigma \in S_p} a_{\ell_{\sigma(1)} i_1} \cdots a_{\ell_{\sigma(p)} i_p} (e_{t_1} \wedge \cdots \wedge e_{t_p}) \\ &= \sum_{\sigma \in S_p} \operatorname{sgn}(\sigma) a_{\ell_{\sigma(1)} i_1} \cdots a_{\ell_{\sigma(p)} i_p} e_L \\ &= \det(A_{L,H}) e_L. \quad \bullet\end{aligned}$$

Multiplication in the exterior algebra $\bigwedge(V)$ is determined by the products $e_H \wedge e_K$ of pairs of basis elements. Let us introduce the following notation: If $H = t_1 \dots t_p$ and $K = \ell_1 \dots \ell_q$ are disjoint increasing lists, then define $\tau_{H,K}$ to be the permutation that rearranges the list $t_1 \dots t_p, \ell_1 \dots \ell_q$ into an increasing list, denoted by $H * K$. Define

$$\rho_{H,K} = \operatorname{sgn}(\tau_{H,K}).$$

With this notation, Lemma 9.152 says that

$$e_H \wedge e_K = \begin{cases} 0 & \text{if } H \cap K \neq \varnothing \\ \rho_{H,K} e_{H*K} & \text{if } H \cap K = \varnothing. \end{cases}$$

Example 9.159.
If $H = 1, 3, 4$ and $K = 2, 6$ are increasing lists, then

$$H * K = 1, 2, 3, 4, 6,$$

and

$$\tau_{H,K} = \begin{pmatrix} 1 & 3 & 4 & 2 & 6 \\ 1 & 2 & 3 & 4 & 6 \end{pmatrix} = (2\ 4\ 3).$$

Therefore,

$$\rho_{H,K} = \operatorname{sgn} \tau_{H,K} = +1,$$

and

$$e_H \wedge e_K = \big(e_1 \wedge e_3 \wedge e_4\big) \wedge \big(e_2 \wedge e_6\big) = e_1 \wedge e_2 \wedge e_3 \wedge e_4 \wedge e_6 = e_{H*K}. \quad \blacktriangleleft$$

Proposition 9.160. *Let $A = [a_{ij}]$ be an $n \times n$ matrix with entries in a commutative ring k.*

(i) *If $I = i_1, \ldots, i_p$ is an increasing p-list and $x_{i_1}, \ldots, x_{i_p}$ are the corresponding columns of A, then denote $x_{i_1} \wedge \cdots \wedge x_{i_p}$ by x_I. If $J = j_1, \ldots, j_q$ is an increasing q-list, then*

$$x_I \wedge x_J = \sum_{H,K} \rho_{H,K} \det(A_{H,I}) \det(A_{K,J}) e_{H*K},$$

where $H * K$ is the increasing $(p+q)$-list formed from $H \cup K$ when $H \cap K = \varnothing$.

(ii) ***Laplace[23] expansion down the jth column***: *For each fixed j,*

$$\det(A) = (-1)^{1+j} a_{1j} \det(A_{1'j'}) + \cdots + (-1)^{n+j} a_{nj} \det(A_{n'j'}),$$

where $A_{i'j'}$ is the $(n-1) \times (n-1)$ submatrix obtained from A by deleting its ith row and jth column.

(iii) ***Laplace expansion across the ith row***: *For each fixed i,*

$$\det(A) = (-1)^{i+1} a_{i1} \det(A_{i'1'}) + \cdots + (-1)^{i+n} a_{in} \det(A_{i'n'}).$$

Proof. (i) By the lemma,

$$\begin{aligned} x_I \wedge x_J &= \sum_H \det(A_{H,I}) e_H \wedge \sum_K \det(A_{K,J}) e_K \\ &= \sum_{H,K} \det(A_{H,I}) e_H \wedge \det(A_{K,J}) e_K \\ &= \sum_{H,K} \det(A_{H,I}) \det(A_{K,J}) e_H \wedge e_K \\ &= \sum_{H,K} \rho_{H,K} \det(A_{H,I}) \det(A_{K,J}) e_{H*K}. \end{aligned}$$

(ii) If $I = j$ has only one element, and if $J = j' = 1, \ldots, \widehat{j}, \ldots, n$ is its complement, then

$$\begin{aligned} x_j \wedge x_{j'} &= x_j \wedge x_1 \wedge \cdots \wedge \widehat{x_j} \wedge \cdots \wedge x_n \\ &= (-1)^{j-1} x_1 \wedge \cdots \wedge x_n \\ &= (-1)^{j-1} \det(A) e_1 \wedge \cdots \wedge e_n, \end{aligned}$$

because $j, 1, \ldots, \widehat{j}, \ldots, n$ can be put in increasing order by $j - 1$ transpositions. On the other hand, we can evaluate $x_j \wedge x_{j'}$ using part (i):

$$x_j \wedge x_{j'} = \sum_{H,K} \rho_{H,K} \det(A_{H,j}) \det(A_{K,j'}) e_{H*K}.$$

[23] After P. S. Laplace.

In this sum, H has just one element, say, $H = i$, while K has $n-1$ elements; thus, $K = \ell'$ for some element ℓ. Since $e_h \wedge e_{\ell'} = 0$ if $\{i\} \cap \ell' \neq \varnothing$, we may assume that $i \notin \ell'$; that is, we may assume that $\ell' = i'$. Now, $\det(A_{i,j}) = a_{ij}$ (this is a 1×1 minor), while $\det(A_{K,j'}) = \det(A_{i'j'})$; that is, $A_{i'j'}$ is the submatrix obtained from A by deleting its jth column and its ith row. Hence, if $e_N = e_1 \wedge \cdots \wedge e_n$,

$$\begin{aligned} x_j \wedge x_{j'} &= \sum_{H,K} \rho_{H,K} \det(A_{H,j}) \det(A_{K,j'}) e_{H*K} \\ &= \sum_i \rho_{i,i'} \det(A_{ij}) \det(A_{i'j'}) e_N \\ &= \sum_i (-1)^{i-1} a_{ij} \det(A_{i'j'}) e_N. \end{aligned}$$

Therefore, equating both values for $x_j \wedge x_{j'}$ gives

$$\det(A) = \sum_i (-1)^{i+j} a_{ij} \det(A_{i'j'}),$$

as desired.

(iii) Laplace expansion across the ith row of A is Laplace expansion down the ith column of A^t, and so the result follows because $\det(A^t) = \det(A)$. •

Notice that we have just proved that Laplace expansion across any row or down any column always has the same value; that is, the determinant is independent of the row or column used for the expansion. The Laplace expansions resemble the sums arising in matrix multiplication, and the following matrix was invented to make this resemblance a reality.

Definition. If $A = [a_{ij}]$ is an $n \times n$ matrix with entries in a commutative ring k, then the ***adjoint***[24] of A is the matrix

$$\operatorname{adj}(A) = [C_{ij}],$$

where

$$C_{ij} = (-1)^{i+j} \det(A_{j'i'}).$$

The reversing of indices is deliberate. In words, $\operatorname{adj}(A)$ is the transpose of the matrix whose ij entry is $(-1)^{i+j} \det(A_{i'j'})$. We often call C_{ij} the ***ij-cofactor*** of A.

Corollary 9.161. *If A is an $n \times n$ matrix with entries in a commutative ring k, then*

$$A \operatorname{adj}(A) = \det(A) I = \operatorname{adj}(A) A.$$

[24]There is no connection between the adjoint of a matrix as just defined and the adjoint of a matrix defined on an inner product space.

Proof. Denote the ij entry of $A\,\text{adj}(A)$ by b_{ij}. The definition of matrix multiplication gives

$$b_{ij} = \sum_{p=1}^{n} a_{ip}C_{pj} = \sum_{p=1}^{n} a_{ip}(-1)^{j+p}\det(A_{j'p'}).$$

If $j = i$, then Proposition 9.160 gives

$$b_{ii} = \det(A).$$

If $j \neq i$, consider the matrix M obtained from A by replacing row j with row i. Of course, $\det(M) = 0$, for it has two identical rows. On the other hand, we may compute $\det(M)$ using Laplace expansion across its "new" row j. All the submatrices $M_{j'p'} = A_{j'p'}$, and so all the corresponding cofactors of M and A are equal. The matrix entries of the new row j are a_{ip}, so that

$$0 = \det(M) = (-1)^{i+1}a_{i1}\det(A_{j'1'}) + \cdots + (-1)^{i+n}a_{in}\det(A_{j'n'}).$$

We have shown that $A\,\text{adj}(A)$ is a diagonal matrix having each diagonal entry equal to $\det(A)$.

The proof that $\det(A)I = \text{adj}(A)A$ is similar and it is left to the reader. [We could also adapt the proof of Corollary 3.107, replacing vector spaces by free k-modules, or we could show that $\text{adj}(A^t) = \text{adj}(A)^t$.] •

Definition. An $n \times n$ matrix A with entries in a commutative ring k is ***invertible over*** k if there is a matrix B with entries in k such that

$$AB = I = BA.$$

If k is a field, then invertible matrices are usually called *nonsingular*, and they are characterized by having a nonzero determinant. Consider the matrix with entries in $\mathbb{Z}$:

$$A = \begin{bmatrix} 3 & 1 \\ 1 & 1 \end{bmatrix}.$$

Now $\det(A) = 2 \neq 0$, but it is not invertible over $\mathbb{Z}$. Suppose

$$\begin{bmatrix} 3 & 1 \\ 1 & 1 \end{bmatrix}\begin{bmatrix} a & c \\ b & d \end{bmatrix} = \begin{bmatrix} 3a+b & 3c+d \\ a+b & c+d \end{bmatrix}.$$

If this product is I, then

$$3a + b = 1 = c + d$$
$$3c + d = 0 = a + b.$$

Hence, $b = -a$ and $1 = 3a + b = 2a$; as there is no solution to $1 = 2a$ in $\mathbb{Z}$, the matrix A is not invertible over $\mathbb{Z}$. Of course, A is invertible over $\mathbb{Q}$.

Proposition 9.162. *If k is a commutative ring and $A \in \mathrm{Mat}_n(k)$, then A is invertible if and only if $\det(A)$ is a unit in k.*

Proof. If A is invertible, then there is a matrix B with $AB = I$. Hence,

$$1 = \det(I) = \det(AB) = \det(A)\det(B);$$

this says that $\det(A)$ is a unit in k.

Conversely, assume that $\det(A)$ is a unit in k, so there is an element $u \in k$ with $u\det(A) = 1$. Define

$$B = u\,\mathrm{adj}(A).$$

By Corollary 9.161,

$$AB = Au\,\mathrm{adj}(A) = u\det(A)I = I = u\,\mathrm{adj}(A)A = BA.$$

Thus, A is invertible. •

Here is a proof by exterior algebra of the computation of the determinant of a matrix in block form.

Proposition 9.163. *Let k be a commutative ring, and let*

$$X = \begin{bmatrix} A & C \\ 0 & B \end{bmatrix}$$

be an $(m+n) \times (m+n)$ matrix with entries in k, where A is an $m \times m$ submatrix, and B is an $n \times n$ submatrix. Then

$$\det(X) = \det(A)\det(B).$$

Proof. Let $e_1, \ldots, e_{m+n}$ be the standard basis of k^{m+n}, let $\alpha_1, \ldots, \alpha_m$ be the columns of A (which are also the first m columns of X), and let $\gamma_i + \beta_i$ be the $(m+i)$th column of X, where γ_i is the ith column of C and β_i is the ith column of B.

Now $\gamma_i \in \langle e_1, \ldots, e_m \rangle$, so that $\gamma_i = \sum_{j=1}^m c_{ji} e_j$. Therefore, if $H = 1, 2, \ldots, n$,

$$e_H \wedge \gamma_i = e_H \wedge \sum_{j=1}^m c_{ji} e_j = 0,$$

because each term has a repeated e_j. Using associativity, we see that

$$\begin{aligned} e_H \wedge (\gamma_1 + \beta_1) \wedge (\gamma_2 + \beta_2) \wedge \cdots \wedge (\gamma_n + \beta_n) \\ &= e_H \wedge \beta_1 \wedge (\gamma_2 + \beta_2) \wedge \cdots \wedge (\gamma_n + \beta_n) \\ &= e_H \wedge \beta_1 \wedge \beta_2 \wedge \cdots \wedge (\gamma_n + \beta_n) \\ &= e_H \wedge \beta_1 \wedge \beta_2 \wedge \cdots \wedge \beta_n. \end{aligned}$$

Hence, if $J = m+1, m+2, \ldots, m+n$,

$$\begin{aligned}
\det(X)e_H \wedge e_J &= \alpha_1 \wedge \cdots \wedge \alpha_m \wedge (\gamma_1 + \beta_1) \wedge \cdots \wedge (\gamma_n + \beta_n) \\
&= \det(A)e_H \wedge (\gamma_1 + \beta_1) \wedge \cdots \wedge (\gamma_n + \beta_n) \\
&= \det(A)e_H \wedge \beta_1 \wedge \cdots \wedge \beta_n \\
&= \det(A)e_H \wedge \det(B)e_J \\
&= \det(A)\det(B)e_H \wedge e_J.
\end{aligned}$$

Therefore, $\det(X) = \det(A)\det(B)$. •

Corollary 9.164. *If $A = [a_{ij}]$ is a triangular $n \times n$ matrix, that is, $a_{ij} = 0$ for all $i < j$ (lower triangular) or $a_{ij} = 0$ for all $i > j$ (upper triangular), then*

$$\det(A) = \prod_{i=1}^{n} a_{ii};$$

that is, $\det(A)$ *is the product of the diagonal entries.*

Proof. An easy induction on $n \geq 1$, using Laplace expansion down the first column (for upper triangular matrices) and the proposition for the inductive step. •

Although the definition of determinant of a matrix A in terms of the wedge of its columns gives an obvious algorithm for computing it, there is a more efficient means of calculating $\det(A)$ when its entries lie in a field. Using Gaussian elimination, there are elementary row operations changing A into an upper triangular matrix T:

$$A \to A_1 \to \cdots \to A_r = T.$$

Keep a record of the operations used. For example, if $A \to A_1$ is an operation of Type I, which multiplies a row by a unit c, then $c\det(A) = \det(A_1)$ and so $\det(A) = c^{-1}\det(A_1)$; if $A \to A_1$ is an operation of Type II, which adds a multiple of some row to another one, then $\det(A) = \det(A_1)$; if $A \to A_1$ is an operation of Type III, which interchanges two rows, then $\det(A) = -\det(A_1)$. Thus, the record allows us, eventually, to write $\det(A)$ in terms of $\det(T)$. But since T is upper triangular, $\det(T)$ is the product of its diagonal entries.

Another application of exterior algebra constructs the trace of a map.

Definition. Let k be a commutative ring and let A be a k-algebra. A ***derivation*** of A is a homomorphism $d\colon A \to A$ of k-modules for which

$$d(ab) = (da)b + a(db).$$

In words, a derivation acts like ordinary differentiation in calculus, for we are saying that the product rule, $(fg)' = f'g + fg'$, holds.

Lemma 9.165. *Let k be a commutative ring, and let M be a k-module.*

(i) *If $\varphi\colon M \to M$ is a k-map, then there exists a unique derivation $D_\varphi\colon T(M) \to T(M)$, where $T(M)$ is the tensor algebra on M, which is a graded map (of degree 0) with $D_\varphi|M = \varphi$; that is, for all $p \geq 0$,*

$$D_\varphi\big(T^p(M)\big) \subseteq T^p(M).$$

(ii) *If $\varphi\colon M \to M$ is a k-map, then there exists a unique derivation $d_\varphi\colon \bigwedge(M) \to \bigwedge(M)$, which is a graded map (of degree 0) with $d_\varphi|M = \varphi$; that is, for all $p \geq 0$,*

$$d_\varphi\big(\bigwedge^p(M)\big) \subseteq \bigwedge^p(M).$$

Proof. (i) Define $D_\varphi|k = 1_k$ (recall that $T^0(M) = k$), and define $D_\varphi|T^1(M) = \varphi$ (recall that $T^1(M) = M$). If $p \geq 2$, define $D_\varphi^p\colon T^p(M) \to T^p(M)$ by

$$D_\varphi^p(m_1 \otimes \cdots \otimes m_p) = \sum_{i=1}^{p} m_1 \otimes \cdots \otimes \varphi(m_i) \otimes \cdots \otimes m_p.$$

For each i, the ith summand in the sum is well-defined, because it arises from the k-multilinear function $(m_1, \ldots, m_p) \mapsto m_1 \otimes \cdots \otimes \varphi(m_i) \otimes \cdots \otimes m_p$; it follows that D_φ is well-defined.

It is clear that D_φ is a map of k-modules. To check that D_φ is a derivation, it suffices to consider its action on homogeneous elements $u = u_1 \otimes \cdots \otimes u_p$ and $v = v_1 \otimes \cdots \otimes v_q$.

$$\begin{aligned} D_\varphi(uv) &= D_\varphi(u_1 \otimes \cdots \otimes u_p \otimes v_1 \otimes \cdots \otimes v_q) \\ &= \sum_{i=1}^{p} u_1 \otimes \cdots \otimes \varphi(u_i) \otimes \cdots \otimes u_p \otimes v \\ &\quad + \sum_{j=1}^{q} u \otimes v_1 \otimes \cdots \otimes \varphi(v_j) \otimes \cdots \otimes v_q \\ &= D_\varphi(u)v + uD_\varphi(v) \end{aligned}$$

We leave the proof of uniqueness to the reader.

(ii) Define $d_\varphi\colon \bigwedge(M) \to \bigwedge(M)$ using the same formula as that for D_φ after replacing $\otimes$ by $\wedge$. To see that this is well-defined, we must show that $D_\varphi(J) \subseteq J$, where J is the two-sided ideal generated by all elements of the form $m \otimes m$. It suffices to prove, by induction on $p \geq 2$, that $D_\varphi(J^p) \subseteq J$, where $J^p = J \cap T^p(M)$. The base step $p = 2$ follows from the identity, for $a, b \in M$,

$$a \otimes b + b \otimes a = (a + b) \otimes (a + b) - a \otimes a - b \otimes b \in J.$$

The inductive step follows from the identity, for $a, c \in M$ and $b \in J^{p-1}$,

$$\begin{aligned} a \otimes b \otimes c + J &= -a \otimes c \otimes b + J \\ &= c \otimes a \otimes b + J \\ &= -c \otimes b \otimes a + J \quad \bullet \end{aligned}$$

Proposition 9.166. *Let k be a commutative ring, and let M be a finitely generated free k-module with basis $e_1, \ldots, e_n$. If $\varphi\colon M \to M$ is a k-map and $d_\varphi\colon \bigwedge(M) \to \bigwedge(M)$ is the derivation it determines, then*

$$d_\varphi \big| \bigwedge^n(M) = \operatorname{tr}(\varphi) e_L,$$

where $e_L = e_1 \wedge \cdots \wedge e_n$.

Proof. By Lemma 9.165(ii), we have $d_\varphi\colon \bigwedge^n(M) \to \bigwedge^n(M)$. Since M is a free k-module of rank n, the binomial theorem gives $\bigwedge^n(M) \cong k$. Hence, $d_\varphi(e_L) = ce_L$ for some $c \in k$; we now show that $c = \operatorname{tr}(\varphi)$. Now $\varphi(e_i) = \sum a_{ji} e_j$.

$$\begin{aligned} d_\varphi(e_L) &= \sum_r e_1 \wedge \cdots \wedge \varphi(e_r) \wedge \cdots \wedge e_n \\ &= \sum_r e_1 \wedge \cdots \wedge \sum a_{jr} e_j \wedge \cdots \wedge e_n \\ &= \sum_r e_1 \wedge \cdots \wedge a_{rr} e_r \wedge \cdots \wedge e_n \\ &= \sum_r a_{rr} e_L \\ &= \operatorname{tr}(\varphi) e_L. \quad \bullet \end{aligned}$$

EXERCISES

9.95 Let k be a commutative ring, and let V and W be free k-modules of ranks m and n, respectively.

(i) Prove that if $f\colon V \to V$ is a k-map, then

$$\det(f \otimes 1_W) = [\det(f)]^n.$$

(ii) Prove that if $f\colon V \to V$ and $g\colon W \to W$ are k-maps, then

$$\det(f \otimes g) = [\det(f)]^n [\det(g)]^m.$$

9.96 (i) Let $z_1, \ldots, z_n$ be elements in a commutative ring k, and consider the ***Vandermonde matrix***

$$V(z_1, \ldots, z_n) = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ z_1 & z_2 & \cdots & z_n \\ z_1^2 & z_2^2 & \cdots & z_n^2 \\ \vdots & \vdots & \vdots & \vdots \\ z_1^{n-1} & z_2^{n-1} & \cdots & z_n^{n-1} \end{bmatrix}.$$

Prove that $\det(V(z_1, \ldots, z_n)) = \prod_{i<j}(z_j - z_i)$.

(ii) If $f(x) = \prod_i (x - z_i)$ has discriminant D, prove that $D = \det(V(z_1, \ldots, z_n))$.

(iii) Prove that if $z_1, \ldots, z_n$ are distinct elements of a field k, then $V(z_1, \ldots, z_n)$ is nonsingular.

9.97 Define a ***tridiagonal matrix*** to be an $n \times n$ matrix of the form

$$T[x_1, \ldots, x_n] = \begin{bmatrix} x_1 & 1 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 \\ -1 & x_2 & 1 & 0 & \cdots & 0 & 0 & 0 & 0 \\ 0 & -1 & x_3 & 1 & \cdots & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & x_4 & \cdots & 0 & 0 & 0 & 0 \\ & & \vdots & & \ddots & & \vdots & & \\ 0 & 0 & 0 & 0 & \cdots & x_{n-3} & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & \cdots & -1 & x_{n-2} & 1 & 0 \\ 0 & 0 & 0 & 0 & \cdots & 0 & -1 & x_{n-1} & 1 \\ 0 & 0 & 0 & 0 & \cdots & 0 & 0 & -1 & x_n \end{bmatrix}.$$

(i) If $D_n = \det(T[x_1, \ldots, x_n])$, prove that $D_1 = x_1$, $D_2 = x_1x_2 + 1$, and, for all $n > 2$,

$$D_n = x_n D_{n-1} + D_{n-2}.$$

(ii) Prove that if all $x_i = 1$, then $D_n = F_{n+1}$, the nth Fibonacci number. (Recall that $F_0 = 0$, $F_1 = 1$, and $F_n = F_{n-1} + F_{n-2}$ for all $n \geq 2$.)

9.98 If a matrix A is a direct sum of square blocks,

$$A = B_1 \oplus \cdots \oplus B_t,$$

prove that $\det(A) = \prod_i \det(B_i)$.

9.99 If A and B are $n \times n$ matrices with entries in a commutative ring R, prove that AB and BA have the same characteristic polynomial.

Hint. (***Goodwillie***)

$$\begin{bmatrix} I & B \\ 0 & I \end{bmatrix} \begin{bmatrix} 0 & 0 \\ A & AB \end{bmatrix} \begin{bmatrix} I & -B \\ 0 & I \end{bmatrix} = \begin{bmatrix} BA & 0 \\ A & 0 \end{bmatrix}.$$

9.10 Lie Algebras

There are interesting examples of nonassociative algebras, the most important of which are the Lie algebras. In the late nineteenth century, Sophus Lie (pronounced LEE) studied the solution space S of a system of partial differential equations using a group G of

transformations of S. The underlying set of G is a differentiable manifold and the group operation is a C^∞-function; such groups are called ***Lie groups***. The solution space is intimately related to its Lie group G; in turn, G is studied using its *Lie algebra*, a considerably simpler object, which arises as the tangent space at the identity element of G. Aside from this fundamental reason for their study, Lie algebras turn out to be the appropriate way to deal with families of linear transformations on a vector space (in contrast to the study of canonical forms of a single linear transformation given in the first sections of this chapter). Moreover, the classification of the simple finite-dimensional complex Lie algebras, due to W. Killing and E. Cartan at the turn of the twentieth century, served as a model for the recent classification of all finite simple groups. It was C. Chevalley who recognized that one could construct analogous families of finite simple groups by imitating the construction of simple Lie algebras.

Before giving the definition of a Lie algebra, let us first present an allied definition. We have already defined derivations of rings; let us now generalize the notion a bit.

Definition. Let k be a commutative ring. A ***not necessarily associative k-algebra*** A is a k-module equipped with some multiplication $A \times A \to A$, denoted by $(a, b) \mapsto ab$, such that

(i) $a(b + c) = ab + ac$ and $(b + c)a = ba + ca$ for all $a, b, c \in A$;

(ii) $ua = au$ for all $u \in k$ and $a \in A$;

(iii) $a(ub) = (au)b = u(ab)$ for all $u \in k$ and $a, b \in A$.

A ***derivation*** of A is a k-map $d \colon A \to A$ for which

$$d(ab) = (da)b + a(db).$$

Aside from ordinary differentiation in calculus, which is a derivation because the product rule holds, $(fg)' = f'g + fg'$, another example is provided by the $\mathbb{R}$-algebra A of all real valued functions $f(x_1, \ldots, x_n)$ of several variables. The partial derivatives $\partial/\partial x_i$ are derivations, for $i = 1, \ldots, n$.

The composite of two derivations need not be a derivation. For example, if $d \colon A \to A$ is a derivation, then $d^2 = d \circ d \colon A \to A$ satisfies the equation

$$d^2(fg) = d^2(f)g + 2d(f)d(g) + fd^2(g);$$

the mixed term $2d(f)d(g)$ is the obstruction to d^2 being a derivation. More generally, we may generalize the Leibniz formula (Exercise 1.6 on page 12) from ordinary differentiation on the ring of all C^∞-functions to a derivation on any not necessarily associative algebra A. If $f, g \in A$, then

$$d^n(fg) = \sum_{i=0}^{n} \binom{n}{i} d^i f \cdot d^{n-i} g.$$

It is still worthwhile to compute the composite of two derivations d_1 and d_2. If A is a not necessarily associative algebra and $f, g \in A$, then

$$\begin{aligned} d_1 d_2(fg) &= d_1 [(d_2 f)g + f(d_2 g)] \\ &= (d_1 d_2 f)g + (d_2 f)(d_1 g) + (d_1 f)(d_2 g) + f(d_1 d_2 g). \end{aligned}$$

Of course,

$$d_2 d_1(fg) = (d_2 d_1 f)g + (d_1 f)(d_2 g) + (d_2 f)(d_1 g) + f(d_2 d_1 g).$$

If we denote $d_1 d_2 - d_2 d_1$ by $[d_1, d_2]$, then subtraction gives

$$[d_1, d_2](fg) = ([d_1, d_2] f)g + f([d_1, d_2]g);$$

that is, $[d_1, d_2] = d_1 d_2 - d_2 d_1$ is a derivation.

Example 9.167.
If k is a commutative ring, equip $\mathrm{Mat}_n(k)$ with the ***bracket operation***:

$$[A, B] = AB - BA.$$

Of course, A and B commute if and only if $[A, B] = 0$. It is easy to find examples showing that the bracket operation is not associative. However, for any fixed $n \times n$ matrix M, the function

$$\mathrm{ad}_M \colon \mathrm{Mat}_n(k) \to \mathrm{Mat}_n(k),$$

defined by

$$\mathrm{ad}_M \colon A \mapsto [M, A],$$

is a derivation:

$$[M, [A, B]] = [[M, A], B] + [A, [M, B]].$$

The verification of this identity should be done once in one's life. ◂

The definition of Lie algebra involves a vector space with a multiplication generalizing the "bracket."

Definition. If k is a field, then a ***Lie algebra*** over k is a vector space L over k equipped with a bilinear operation $L \times L \to L$, denoted by $(a, b) \mapsto [a, b]$ (and called ***bracket***), such that

(i) $[a, a] = 0$ for all $a \in L$;

(ii) For each $a \in L$, the function $\mathrm{ad}_a \colon b \mapsto [a, b]$ is a derivation.

For all $u, v \in L$, bilinearity gives

$$[u + v, u + v] = [u, u] + [u, v] + [v, u] + [v, v],$$

which, when coupled with the first axiom $[a, a] = 0$, gives

$$[u, v] = -[v, u];$$

that is, bracket is ***anticommutative***. The second axiom is often written out in more detail. If $b, c \in L$, then their product in L is denoted by $[b, c]$; that ad_a is a derivation is to say

$$[a, [b, c]] = [[a, b], c] + [b, [a, c]];$$

rewriting,

$$[a, [b, c]] - [b, [a, c]] - [[a, b], c] = 0.$$

The anticommutativity from the first axiom now gives the ***Jacobi identity***:

$$[a, [b, c]] + [b, [c, a]] + [c, [a, b]] = 0 \quad \text{for all } a, b, c \in L.$$

Thus, a vector space L is a Lie algebra if and only if $[a, a] = 0$ for all $a \in L$ and the Jacobi identity holds.

Here are some examples of Lie algebras.

Example 9.168.

(i) If V is a vector space over a field k, define $[a, b] = 0$ for all $a, b \in V$. It is obvious that V so equipped is a Lie algebra, and it is called an ***abelian*** Lie algebra.

(ii) In $\mathbb{R}^3$, define $[u, v] = u \times v$, the vector product (or cross product) defined in calculus. It is routine to check that $v \times v = 0$ and that the Jacobi identity holds, so that $\mathbb{R}^3$ is a Lie algebra. This example may be generalized: For every field k, cross product can be defined on the vector space k^3 making it a Lie algebra.

(iii) A ***subalgebra*** S of a Lie algebra L over a field k is a subspace that is closed under bracket: If $a, b \in S$, then $[a, b] \in S$. It is easy to see that every subalgebra is itself a Lie algebra.

(iv) If k is a field, then $\text{Mat}_n(k)$ is a Lie algebra if we define bracket by

$$[A, B] = AB - BA.$$

We usually denote this Lie algebra by $\mathfrak{gl}(n, k)$. This example is quite general, for it is a theorem of I. D. Ado that every finite-dimensional Lie algebra over a field k of characteristic 0 is isomorphic to a subalgebra of $\mathfrak{gl}(n, k)$ for some n (see Jacobson, *Lie Algebras*, page 202).

(v) An interesting subalgebra of $\mathfrak{gl}(n, k)$ is $\mathfrak{sl}(n, k)$, which consists of all $n \times n$ matrices of trace 0. In fact, if G is a Lie group whose associated Lie algebra is $\mathfrak{g}$, then there is an analog of exponentiation $\mathfrak{g} \to G$. In particular, if $\mathfrak{g} = \mathfrak{gl}(n, \mathbb{C})$, then this map is

exponentiation $A \mapsto e^A$. Thus, Proposition 9.52(viii) shows that exponentiation sends $\mathfrak{sl}(n, \mathbb{C})$ into $\mathrm{SL}(n, \mathbb{C})$.

(vi) If A is any algebra over a field k, then

$$\mathfrak{Der}(A/k) = \{\text{all derivations } d \colon A \to A\},$$

with bracket $[d_1, d_2] = d_1 d_2 - d_2 d_1$, is a Lie algebra.

It follows from the Leibniz rule that if k has characteristic $p > 0$, then d^p is a derivation for every $d \in \mathfrak{Der}(A/k)$, for $\binom{p}{i} \equiv 0 \bmod p$ whenever $0 < i < p$. (This is an example of what is called a ***restricted Lie algebra*** of characteristic p.)

There is a Galois theory for certain purely inseparable extensions, due to N. Jacobson (see Jacobson, *Basic Algebra* II, pages 533–536). If k is a field of characteristic $p > 0$ and E/k is a finite purely inseparable extension of *height* 1, that is, $\alpha^p \in k$, for all $\alpha \in E$, then there is a bijection between the family of all intermediate fields and the restricted Lie subalgebras of $\mathfrak{Der}(E/k)$, given by

$$B \mapsto \mathfrak{Der}(E/B);$$

the inverse of this function is given by

$$\mathfrak{L} \mapsto \big\{e \in E : D(e) = 0 \text{ for all } D \in \mathfrak{L}\big\}. \quad ◀$$

Not surprisingly, all Lie algebras over a field k form a category.

Definition. If L and L' are Lie algebras over a field k, then a function $f \colon L \to L'$ is a ***Lie homomorphism*** if f is a k-linear transformation that preserves bracket: For all $a, b \in L$,

$$f([a, b]) = [fa, fb].$$

Definition. An ***ideal*** of a Lie algebra L is a subspace I such that $[x, a] \in I$ for every $x \in L$ and $a \in I$.

Even though a Lie algebra need not be commutative, its anticommutativity shows that every left ideal (as just defined) is necessarily a right ideal; that is, every ideal is two-sided.

A Lie algebra L is called ***simple*** if $L \neq \{0\}$ and L has no nonzero proper ideals.

Definition. If I is an ideal in L, then the ***quotient*** L/I is the quotient space (considering L as a vector space and I as a subspace) with bracket defined by

$$[a + I, b + I] = [a, b] + I.$$

It is easy to check that this bracket on L/I is well-defined. If $a' + I = a + I$ and $b' + I = b + I$, then $a - a' \in I$ and $b - b' \in I$, and so

$$\begin{aligned}[a', b'] - [a, b] &= [a', b'] - [a', b] + [a', b] - [a, b] \\ &= [a', b' - b] + [a' - a, b'] \in I.\end{aligned}$$

Example 9.169.

(i) If $f : L \to L'$ is a Lie homomorphism, then its ***kernel*** is defined as usual:

$$\ker f = \{a \in L : f(a) = 0\}.$$

It is easy to see that $\ker f$ is an ideal in L.

Conversely, the ***natural map*** $\nu : L \to L/I$, defined by $a \mapsto a + I$, is a Lie homomorphism whose kernel is I. Thus, a subspace of L is an ideal if and only if it is the kernel of some Lie homomorphism.

(ii) If I and J are ideals in a Lie algebra L, then

$$IJ = \Big\{\sum_r [i_r, j_r] : i_r \in I \text{ and } j_r \in J\Big\}.$$

In particular, $L^2 = LL$ is the analog for Lie algebras of the commutator subgroup of a group: $L^2 = \{0\}$ if and only if L is abelian.

(iii) There is an analog for Lie algebras of the derived series of a group. The ***derived series*** of a Lie algebra L is defined inductively:

$$L^{(0)} = L; \quad L^{(n+1)} = (L^{(n)})^2.$$

A Lie algebra L is called ***solvable*** if there is some $n \geq 0$ with $L^{(n)} = \{0\}$.

(iv) There is an analog for Lie algebras of the descending central series of a group. The ***descending central series*** is defined inductively:

$$L_1 = L; \quad L_{n+1} = LL_n.$$

A Lie algebra L is called ***nilpotent*** if there is some $n \geq 0$ with $L_n = \{0\}$. ◂

We merely mention the first two theorems in the subject. If L is a Lie algebra and $a \in L$, then $\mathrm{ad}_a : L \to L$, given by $\mathrm{ad}_a : x \mapsto [a, x]$, is a linear transformation on L (viewed merely as a vector space). We say that a is ***ad-nilpotent*** if ad_a is a nilpotent operator; that is, $(\mathrm{ad}_a)^m = 0$ for some $m \geq 1$.

Theorem (Engel's Theorem).

(i) *If L is a finite-dimensional Lie algebra over any field k, then L is nilpotent if and only if every $a \in L$ is* ad-*nilpotent.*

(ii) *If L is a Lie subalgebra of $\mathfrak{gl}(n, k)$ all of whose elements A are nilpotent matrices, then L can be put into strict upper triangular form* (*all diagonal entries are* 0); *that is, there is a nonsingular matrix P so that PAP^{-1} is strictly upper triangular for every $A \in L$.*

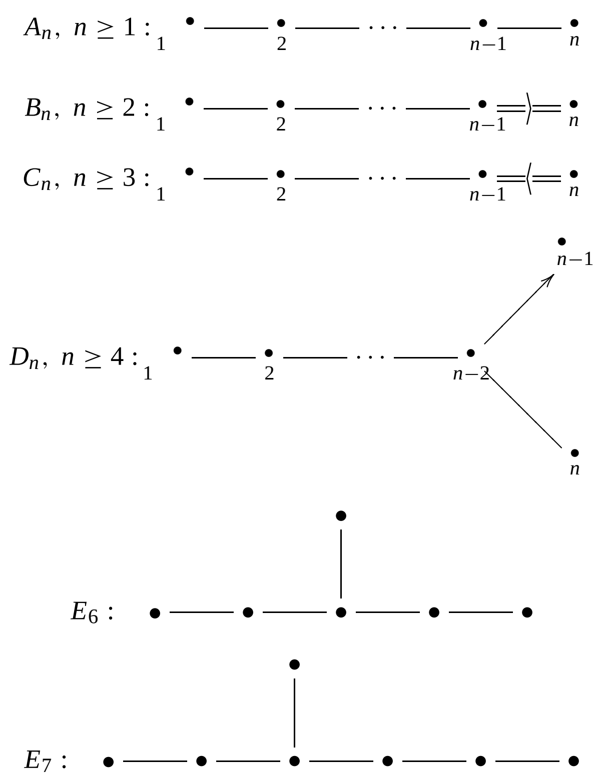
Proof. See Humphreys, *Introduction to Lie Algebras and Representation Theory*, page 12. •

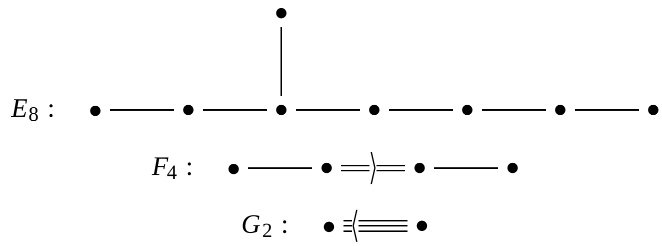
Compare Engel's theorem with Exercise 9.43(i) on page 682, which is the much simpler version for a single nilpotent matrix. Nilpotent Lie algebras are so called because of Engel's theorem; it is likely that nilpotent groups are so called by analogy. Corollary 5.48, which states that every finite p-group can be imbedded as a subgroup of unitriangular matrices over $\mathbb{F}_p$, may be viewed as a group-theoretic analog of Engel's theorem.

Theorem (Lie's Theorem). *Every solvable subalgebra L of $\mathfrak{gl}(n, k)$, where k is an algebraically closed field, can be put into* (*not necessarily strict*) *upper triangular form; that is, there is a nonsingular matrix P so that PAP^{-1} is upper triangular for every $A \in L$.*

Proof. See Humphreys, *Introduction to Lie Algebras and Representation Theory*, page 16. •

Further study of Lie algebras leads to the classification of all finite-dimensional simple Lie algebras, due to E. Cartan and W. Killing, over an algebraically closed field of characteristic 0 (recently, the classification of all finite-dimensional simple Lie algebras in characteristic p has been given, where $p > 7$). To each such algebra, they associated a certain geometric configuration called a *root system*, which is characterized by a *Cartan matrix*. Cartan matrices are, in turn, characterized by ***Dynkin diagrams***.





Every Dynkin diagram arises from a simple Lie algebra over $\mathbb{C}$, and two such algebras are isomorphic if and only if they have the same Dynkin diagram. We refer the reader to Humphreys, *Introduction to Lie Algebras and Representation Theory*, Chapter IV, and Jacobson, *Lie Algebras*, Chapter IV.

There are other not necessarily associative algebras of interest. ***Jordan algebras*** are commutative algebras A in which the Jacobi identity is replaced by

$$(x^2y)x = x^2(yx)$$

for all $x, y \in A$. They were introduced by P. Jordan to provide an algebraic setting for doing quantum mechanics. An example of a Jordan algebra is a subspace of all $n \times n$ matrices, over a field of characteristic not 2, equipped with the binary operation $A * B$, where

$$A * B = \tfrac{1}{2}(AB + BA).$$

Another source of not necessarily associative algebras comes from combinatorics. The usual construction of a projective plane $P(k)$ over a field k, as the family of all lines in k^3 passing through the origin, leads to descriptions of its points by "homogeneous coordinates" $[x, y, z]$, where $x, y, z \in k$. Define an abstract ***projective plane*** to be an ordered pair $(X, \mathcal{L})$, where X is a finite set and $\mathcal{L}$ is a family of subsets of X, called *lines*, subject to the following axioms:

(i) All lines have the same number of points;

(ii) Given any two points in X, there is a unique line containing them.

We want to introduce homogeneous coordinates to describe the points of such a projective plane, but there is no field k given at the outset. Instead, we look at a collection $\mathcal{K}$ of functions on X, called *collineations*, and we equip $\mathcal{K}$ with two binary operations (called addition and multiplication). In general, $\mathcal{K}$ is a not necessarily associative algebra, but certain of its algebraic properties—commutativity and associativity of multiplication—correspond to geometric properties of the projective plane—a theorem of Pappus and a theorem of Desargues, respectively.

An interesting nonassociative algebra is the ***Cayley numbers*** (sometimes called *octonions*), which is an eight-dimensional real vector space containing the quaternions as a subalgebra (see the article by Curtis in Albert, *Studies in Modern Algebra*). Indeed, Cayley numbers form a real division not necessarily associative algebra in the sense that every nonzero element has a multiplicative inverse. The Cayley numbers acquire added interest

(as do other not necessarily associative algebras) because its automorphism group has interesting properties. For example, the exceptional simple Lie algebra E_8 is isomorphic to the Lie algebra of all the derivations of the Cayley numbers, while the Monster, the largest sporadic finite simple group, is the automorphism group of a certain nonassociative algebra constructed by R. Griess.

EXERCISES

9.100 Consider the de Rham complex when $n = 2$:

$$0 \to \Omega^0(X) \xrightarrow{d^0} \Omega^1(X) \xrightarrow{d^1} \Omega^2(X) \to 0.$$

Prove that if $f(x, y) \in A(X) = \Omega^0(X)$, then

$$d^0 f = \frac{\partial f}{\partial x}\, dx + \frac{\partial f}{\partial y}\, dy,$$

and that if $P dx + Q dy$ is a 1-form, then

$$d^1(P dx + Q dy) = \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y}\right) dx \wedge dy.$$

9.101 Prove that if L and L' are nonabelian two-dimensional Lie algebras, then $L \cong L'$.

9.102 (i) Prove that the ***center*** of a Lie algebra L, defined by

$$Z(L) = \{a \in L : [a, x] = 0 \text{ for all } x \in L\},$$

is an abelian ideal in L.

(ii) Give an example of a Lie algebra L for which $Z(L) = \{0\}$.

(iii) If L is nilpotent and $L \neq \{0\}$, prove that $Z(L) \neq \{0\}$.

9.103 Prove that if L is an n-dimensional Lie algebra, then $Z(L)$ cannot have dimension $n - 1$. (Compare Exercise 2.69 on page 95.)

9.104 Equip $\mathbb{C}^3$ with a cross product (using the same formula as the cross product on $\mathbb{R}^3$). Prove that

$$\mathbb{C}^3 \cong \mathfrak{sl}(2, \mathbb{C}).$$

10

Homology

10.1 INTRODUCTION

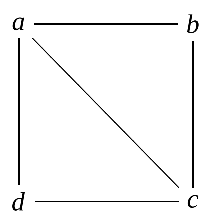
When I was a graduate student, homological algebra was an unpopular subject. The general attitude was that it was a grotesque formalism, boring to learn, and not very useful once one had learned it. Perhaps an algebraic topologist was forced to know this stuff, but surely no one else should waste time on it. The few true believers were viewed as workers at the fringe of mathematics who kept tinkering with their elaborate machine, smoothing out rough patches here and there.

This attitude changed dramatically when J.-P. Serre characterized regular local rings using homological algebra (they are the commutative noetherian local rings of "finite global dimension"), for this enabled him to prove that any localization of a regular local ring is itself regular (until then, only special cases of this were known). At the same time, M. Auslander and D. A. Buchsbaum completed work of M. Nagata by using global dimension to prove that every regular local ring is a UFD.

In spite of its newfound popularity, homological algebra still "got no respect." For example, the two theorems just mentioned used the notion of the global dimension of a ring which, in turn, is defined in terms of the *homological dimension* of a module. At that time, I. Kaplansky offered a course in homological algebra. One of his students, S. Schanuel, noticed that there is an elegant relation between different projective resolutions of the same module (see Proposition 7.60). Kaplansky seized this result, nowadays called Schanuel's lemma, for it allowed him to define the homological dimension of a module without having first to develop the fundamental constructs Ext and Tor of homological algebra, and he was then able to prove the theorems of Serre and of Auslander–Buchsbaum (Kaplansky's account of this course can be found in his book, *Commutative Algebra*). However, as more applications were found and as more homology and cohomology theories were invented to solve outstanding problems, resistance to homological algebra waned. Today, it is just

another standard tool in a mathematician's kit.

The basic idea of homology comes from Green's theorem, where a double integral over a region R with holes in it is equal to a line integral on the boundary of R. H. Poincaré recognized that whether a topological space X has different kinds of holes is a kind of connectivity. To illustrate, let us assume that X can be "triangulated;" that is, X can be partitioned into finitely many *n-simplexes*, where $n \geq 0$: points are 0-simplexes, edges are 1-simplexes, triangles are 2-simplexes, tetrahedra are 3-simplexes, and there are higher-dimensional analogs. The question to ask is whether a union of n-simplexes in X that "ought" to be the boundary of some $(n+1)$-simplex actually is such a boundary. For example, when $n = 0$, two points a and b in X ought to be the boundary (endpoints) of a path in X; if there is a path in X joining all points a and b, then X is called *path connected*; if there is no such path, then X has a 0-dimensional hole. For an example of a one-dimensional hole, let X be the *punctured plane*; that is, the plane with the origin deleted. The perimeter of a triangle Δ ought to be the boundary of a 2-simplex, but this is not so if Δ contains the origin in its interior; thus, X has a one-dimensional hole. If X were missing a line segment containing the origin, or even a small disk containing the origin, this hole would still be one-dimensional; we are not considering the size of the hole, but the size of the possible boundary. We must keep our eye on the doughnut and not upon the hole!



For example, in the rectangle drawn above, consider the triangle $[a, b, c]$ with vertices a, b, c and edges $[a, b]$, $[b, c]$, $[a, c]$. Its boundary $\partial[a, b, c]$ should be $[a, b]+[b, c]+[c, a]$. But edges are oriented (think of $[a, c]$ as a path from a to c and $[c, a]$ as the reverse path from c to a), so let us write $[c, a] = -[a, c]$. Thus, the boundary is

$$\partial[a, b, c] = [a, b] - [a, c] + [b, c].$$

Similarly, let us define the boundary of $[a, b]$ to be its endpoints:

$$\partial[a, b] = b - a.$$

We note that

$$\begin{aligned}\partial(\partial[a, b, c]) &= \partial([a, b] - [a, c] + [b, c]) \\ &= b - a - (c - a) + c - b \\ &= 0.\end{aligned}$$

The rectangle with vertices a, b, c, d is the union of two triangles $[a, b, c] + [a, c, d]$, and we check that its boundary is $\partial[a, b, c] + \partial[a, c, d]$ (note that the diagonal $[a, c]$ occurs

twice, with different signs, and so it cancels, as it should). We see that the formalism suggests the use of signs to describe boundaries as certain linear combinations u of edges or points for which $\partial(u) = 0$.

Such ideas lead to the following construction. For each $n \geq 0$, consider all formal linear combinations of n-simplexes; that is, form the free abelian group $C_n(X)$ with basis all n-simplexes, and call such linear combinations *n-chains*. Some of these n-chains ought to be boundaries of some union of $(n+1)$-simplexes; call them *n-cycles* (for example, adding the three edges of a triangle, with appropriate choice of signs, is a 1-cycle). Certain n-chains actually are boundaries, and these are called *n-boundaries* (if Δ is a triangle in the punctured plane X, not having the origin in its interior, then the alternating sum of the edges of Δ is a 1-boundary; on the other hand, if the origin does lie in the interior of Δ, then the alternating sum is a 1-cycle but not a 1-boundary). The family of all the n-cycles, $Z_n(X)$, and the family of all the n-boundaries, $B_n(X)$, are subgroups of $C_n(X)$. A key ingredient in the construction of homology groups is that the subgroups Z_n and B_n can be defined in terms of homomorphisms: there are *boundary homomorphisms* $\partial_n : C_n(X) \to C_{n-1}(X)$ with $Z_n = \ker \partial_n$ and $B_n = \operatorname{im} \partial_{n+1}$, and so there is a sequence of abelian groups and homomorphisms

$$\cdots \to C_3(X) \xrightarrow{\partial_3} C_2(X) \xrightarrow{\partial_2} C_1(X) \xrightarrow{\partial_1} C_0(X).$$

It turns out, for all $n \geq 1$, that $\partial_n \partial_{n+1} = 0$, from which it follows that

$$B_n(X) \subseteq Z_n(X).$$

The interesting group is the quotient group $Z_n(X)/B_n(X)$, denoted by $H_n(X)$ and called the nth *homology*[1] *group* of X. What survives in this quotient group are the n-dimensional holes; that is, those n-cycles that are not n-boundaries. For example, $H_0(X) = 0$ means that X is path connected: if there are two points $a, b \in X$ that are not connected by a path, then $a - b$ is a cycle that is not a boundary, and so the coset $a - b + B_0(X)$ is a nonzero element of $H_0(X)$. For $n \geq 1$, these groups measure more subtle kinds of connectivity. Topologists modify this construction in two ways. They introduce homology with *coefficients* in an abelian group G by tensoring the sequence of chain groups by G and then taking homology groups; they also consider *cohomology with coefficients* in G by applying the contravariant functor $\operatorname{Hom}(\ , G)$ to the sequence of chain groups and then taking homology groups. Homological algebra arose in trying to compute and to find relations between homology groups of spaces.

[1] I have not been able to discover the etymology of the mathematical term *homology* as used in this context. The word "homology" comes from *homo* + *logos*, and it means "corresponding." Its first usage as a mathematical term occurred in projective geometry in the early 19th century, as the name of a specific type of collineation. The earliest occurrence I have found for its usage in the sense of cycles and boundaries is in an article of H. Poincaré: *Analysis Situs*, Journal de l'École Polytechnique, Series II, First issue, 1895 (and Oeuvres, vol. 5), but he does not explain why he chose the term. Emili Bifet has written, in a private communication, "Consider the projective homology, between two distinct (hyper)planes, given by projection from an exterior point. This homology suggests (and provides) a natural way of deforming the boundary of a simplex contained in one plane into the boundary of the corresponding simplex on the other one. Moreover, it suggests a natural way of deforming a boundary into a point. This could be what Poincaré had in mind."

We have already seen, in Proposition 7.51, that every left R-module M, where R is a ring, has a description by generators and relations. There is an exact sequence

$$0 \to \ker\varphi \xrightarrow{\iota} F \xrightarrow{\varphi} M \to 0,$$

where F is a free left R-module and ι is the inclusion. If R is a PID, then $\ker\varphi$ is free, because every submodule of a free module is itself free; if R is not a PID, then $\ker\varphi$ may not be free. Now take generators and relations of $\ker\varphi$: There is a free module F_1 and an exact sequence

$$0 \to \ker\psi \xrightarrow{\kappa} F_1 \xrightarrow{\psi} \ker\varphi \to 0.$$

If we define $F_1 \to F$ to be the composite $\iota\psi$, then there is a second exact sequence

$$F_1 \xrightarrow{\iota\psi} F \xrightarrow{\varphi} M \to 0,$$

and, iterating this construction, there is a long exact sequence

$$\cdots \to F_3 \to F_2 \to F_1 \to F \to M \to 0.$$

We can view the submodules $\ker(F_n \to F_{n-1})$ as "relations on relations" (nineteenth century algebraists called these higher relations *syzygies*). This long exact sequence resembles the sequence of chain groups in topology. There are other contexts in which such exact sequences exist; many algebraic structures give rise to a sequence of homology groups, and these can be used to translate older theorems into the language of homology. Examples of such theorems are Hilbert's Theorem 90 about algebras (see Corollary 10.129), Whitehead's lemmas about Lie algebras (see Jacobson, *Lie Algebras*, pages 77 and 89), and Theorem 10.22, the Schur–Zassenhaus lemma, about groups. There are methods to compute homology and cohomology groups, and this is the most important contribution of homological algebra to this circle of ideas. Although we can calculate many things without them, the most powerful method of computing homology groups uses *spectral sequences*. When I was a graduate student, I always wanted to be able to say, nonchalantly, that such and such is true "by the usual spectral sequence argument," but I never had the nerve.[2] We will sketch what spectral sequences are at the end of this chapter.

10.2 Semidirect Products

We begin by investigating a basic problem in group theory. A group G having a normal subgroup K can be "factored" into K and G/K; the study of extensions involves the inverse question: How much of G can be recovered from a normal subgroup K and the quotient $Q = G/K$? For example, we know that $|G| = |K||Q|$ if K and Q are finite.

[2] This introduction is adapted from a review I wrote that appeared in *Bulletin of the American Mathematical Society*, Vol. 33, pp. 473–475, 1996; it is reproduced by permission of the American Mathematical Society.

Exactness of a sequence of nonabelian groups,

$$\cdots \to G_{n+1} \xrightarrow{d_{n+1}} G_n \xrightarrow{d_n} G_{n-1} \to \cdots,$$

is defined just as it is for abelian groups: $\operatorname{im} d_{n+1} = \ker d_n$ for all n. Of course, each $\ker d_n$ is a normal subgroup of G_n.

Definition. If K and Q are groups, then an ***extension*** of K by Q is a short exact sequence

$$1 \to K \xrightarrow{i} G \xrightarrow{p} Q \to 1.$$

The notation K is to remind us of kernel, and the notation Q is to remind us of quotient.

There is an alternative usage of the term *extension*, which calls the (middle) group G (not the short exact sequence) an *extension* if it contains a normal subgroup K_1 with $K_1 \cong K$ and $G/K_1 \cong Q$. As do most people, we will use the term in both senses.

Example 10.1.
(i) The direct product $K \times Q$ is an extension of K by Q; it is also an extension of Q by K.

(ii) Both S_3 and $\mathbb{I}_6$ are extensions of $\mathbb{I}_3$ by $\mathbb{I}_2$. On the other hand, $\mathbb{I}_6$ is an extension of $\mathbb{I}_2$ by $\mathbb{I}_3$, but S_3 is not, for S_3 contains no normal subgroup of order 2. ◂

We have just seen, for any given ordered pair of groups, that there always exists an extension of one by the other (their direct product), but there may be other extensions as well. The ***extension problem*** is to classify all possible extensions of a given pair of groups K and Q.

In Chapter 5, on page 283, we discussed the relation between the extension problem and the Jordan–Hölder theorem. If a group G has a composition series

$$G = K_0 \geq K_1 \geq K_2 \geq \cdots \geq K_{n-1} \geq K_n = \{1\}$$

with simple factor groups $Q_1, \ldots, Q_n$, where $Q_i = K_{i-1}/K_i$ for all $i \geq 1$, then G could be recaptured from $Q_n, Q_{n-1}, \ldots, Q_1$ by solving the extension problem n times. Now all finite simple groups have been classified, and so we could survey all finite groups if we could solve the extension problem.

Let us begin by recalling the partition of a group into the cosets of a subgroup. We have already defined a *transversal* of a subgroup K of a group G as a subset T of G consisting of exactly one element from each coset[3] Kt of K.

Definition. If

$$1 \to K \to G \xrightarrow{p} Q \to 1$$

is an extension, then a ***lifting*** is a function $\ell \colon Q \to G$, not necessarily a homomorphism, with $p\ell = 1_Q$.

[3]We have been working with *left* cosets tK, but, in this chapter, the subgroup K will be a normal subgroup, in which case $tK = Kt$ for all $t \in G$. Thus, using right cosets or left cosets is only a matter of convenience.

Given a transversal, we can construct a lifting. For each $x \in Q$, surjectivity of p provides $\ell(x) \in G$ with $p\ell(x) = x$; thus, the function $x \mapsto \ell(x)$ is a lifting. Conversely, given a lifting, we claim that $\operatorname{im}\ell$ is a transversal of K. If Kg is a coset, then $p(g) \in Q$; say, $p(g) = x$. Then $p(g\ell(x)^{-1}) = 1$, so that $a = g\ell(x)^{-1} \in K$ and $Kg = K\ell(x)$. Thus, every coset has a representative in $\ell(Q)$. Finally, we must show that $\ell(Q)$ does not contain two elements in the same coset. If $K\ell(x) = K\ell(y)$, then there is $a \in K$ with $a\ell(x) = \ell(y)$. Apply p to this equation; since $p(a) = 1$, we have $x = y$ and so $\ell(x) = \ell(y)$.

The following group will arise in our discussion of extensions.

Definition. Recall that an *automorphism* of a group K is an isomorphism $K \to K$. The ***automorphism group***, denoted by $\operatorname{Aut}(K)$, is the group of all the automorphisms of K with composition as operation.

Of course, extensions are defined for arbitrary groups K, but we are going to restrict our attention to the special case when K is abelian. If G is an extension of K by Q, it would be confusing to write G multiplicatively and its subgroup K additively. Hence, we shall use the following notational convention: Even though G may not be abelian, additive notation will be used for the operation in G. Corollary 10.4 gives the main reason for this decision.

Proposition 10.2. *Let*

$$0 \to K \xrightarrow{i} G \xrightarrow{p} Q \to 1$$

be an extension of an abelian group K by a group Q, and let $\ell\colon Q \to G$ be a lifting.

(i) *For every $x \in Q$, conjugation $\theta_x\colon K \to K$, defined by*

$$\theta_x\colon a \mapsto \ell(x) + a - \ell(x),$$

is independent of the choice of lifting $\ell(x)$ of x. [*For convenience, we have assumed that i is an inclusion; this merely allows us to write a instead of $i(a)$.*]

(ii) *The function $\theta\colon Q \to \operatorname{Aut}(K)$, defined by $x \mapsto \theta_x$, is a homomorphism.*

Proof. (i) Let us now show that θ_x is independent of the choice of lifting $\ell(x)$ of x. Suppose that $\ell'(x) \in G$ and $p\ell'(x) = x$. There is $b \in K$ with $\ell'(x) = \ell(x) + b$ [for $-\ell(x) + \ell'(x) \in \ker p = \operatorname{im} i = K$]. Therefore,

$$\begin{aligned}\ell'(x) + a - \ell'(x) &= \ell(x) + b + a - b - \ell(x)\\ &= \ell(x) + a - \ell(x),\end{aligned}$$

because K is abelian.

(ii) Now $\theta_x(a) \in K$ because $K \lhd G$, so that each $\theta_x\colon K \to K$; also, θ_x is an automorphism of K, because conjugations are automorphisms.

It remains to prove that $\theta\colon Q \to \mathrm{Aut}(K)$ is a homomorphism. If $x, y \in Q$ and $a \in K$, then

$$\theta_x(\theta_y(a)) = \theta_x(\ell(y) + a - \ell(y)) = \ell(x) + \ell(y) + a - \ell(y) - \ell(x),$$

while

$$\theta_{xy}(a) = \ell(xy) + a - \ell(xy).$$

But $\ell(x) + \ell(y)$ and $\ell(xy)$ are both liftings of xy, so that equality $\theta_x\theta_y = \theta_{xy}$ follows from part (i). •

Roughly speaking, the homomorphism θ tells "how" K is normal in G, for isomorphic copies of a group can sit as normal subgroups of G in different ways. For example, let K be a cyclic group of order 3 and let $Q = \langle x \rangle$ be cyclic of order 2. If $G = K \times Q$, then G is abelian and K lies in the center of G. In this case, $\ell(x) + a - \ell(x) = a$ for all $a \in K$ and $\theta_x = 1_K$. On the other hand, if $G = S_3$, then $K = A_3$ which does not lie in the center; if $\ell(x) = (1\ 2)$, then $(1\ 2)(1\ 2\ 3)(1\ 2) = (1\ 3\ 2)$ and θ_x is not 1_K.

The existence of a homomorphism θ equips K with a scalar multiplication making K a left $\mathbb{Z}Q$-module, where $\mathbb{Z}Q$ is the group ring whose elements are all $\sum_{x \in Q} m_x x$ for $m_x \in \mathbb{Z}$.

Proposition 10.3. *Let K and Q be groups with K abelian. Then a homomorphism $\theta\colon Q \to \mathrm{Aut}(K)$ makes K into a left $\mathbb{Z}Q$-module if scalar multiplication is defined by*

$$xa = \theta_x(a)$$

for all $a \in K$ and $x \in Q$. Conversely, if K is a left $\mathbb{Z}Q$-module, then $x \mapsto \theta_x$ defines a homomorphism $\theta\colon Q \to \mathrm{Aut}(K)$, where $\theta_x\colon a \mapsto xa$.

Proof. Define scalar multiplication as follows. Each $u \in \mathbb{Z}Q$ has a unique expression of the form $u = \sum_{x \in Q} m_x x$, where $m_x \in \mathbb{Z}$ and almost all $m_x = 0$; define

$$\Big(\sum_x m_x x\Big)a = \sum_x m_x \theta_x(a) = \sum_x m_x(xa).$$

We verify the module axioms. Since θ is a homomorphism, $\theta(1) = 1_K$, and so $1a = \theta_1(a)$ for all $a \in K$. That $\theta_x \in \mathrm{Aut}(K)$ implies $x(a + b) = xa + xb$, from which it follows that $u(a + b) = ua + ub$ for all $u \in \mathbb{Z}Q$. Similarly, we check easily that $(u + v)a = ua + va$ for $u, v \in \mathbb{Z}Q$. Finally, $(uv)a = u(va)$ will follow from $(xy)a = x(ya)$ for all $x, y \in Q$; but

$$(xy)a = \theta_{xy}(a) = \theta_x(\theta_y(a)) = \theta_x(ya) = x(ya).$$

The proof of the converse is also routine. •

Corollary 10.4. *If*

$$0 \to K \xrightarrow{i} G \xrightarrow{p} Q \to 1$$

is an extension of an abelian group K by a group Q, then K is a left $\mathbb{Z}Q$-module if we define

$$xa = \ell(x) + a - \ell(x),$$

where $\ell \colon Q \to G$ is a lifting, $x \in Q$, and $a \in K$; moreover, the scalar multiplication is independent of the choice of lifting ℓ.

Proof. Propositions 10.2 and 10.3. •

From now on, we will abbreviate the term "left $\mathbb{Z}Q$-module" to "Q-module."

Recall that a short exact sequence of left R-modules

$$0 \to A \xrightarrow{i} B \xrightarrow{p} C \to 0$$

is *split* if there exists a homomorphism $j \colon C \to B$ with $pj = 1_C$; in this case, the middle module is isomorphic to the direct sum $A \oplus C$. Here is the analogous definition for groups.

Definition. An extension of groups

$$0 \to K \xrightarrow{i} G \xrightarrow{p} Q \to 1$$

is ***split*** if there is a homomorphism $j \colon Q \to G$ with $pj = 1_Q$. The middle group G in a split extension is called a ***semidirect product*** of K by Q.

Thus, an extension is split if and only if there is a lifting, namely, j, that is also a homomorphism. We shall use the following notation: The elements of K shall be denoted by $a, b, c, \ldots$, and the elements of Q shall be denoted by $x, y, z, \ldots$.

Proposition 10.5. *Let G be an additive group having a normal subgroup K.*

(i) *If $0 \to K \xrightarrow{i} G \xrightarrow{p} Q \to 1$ is a split extension, where $j \colon Q \to G$ satisfies $pj = 1_Q$, then $i(K) \cap j(Q) = \{0\}$ and $i(K) + j(Q) = G$.*

(ii) *In this case, each $g \in G$ has a unique expression $g = i(a) + j(x)$, where $a \in K$ and $x \in Q$.*

(iii) *Let K and Q be subgroups of a group G with $K \lhd G$. Then G is a semidirect product of K by Q if and only if $K \cap Q = \{0\}$, $K + Q = G$, and each $g \in G$ has a unique expression $g = a + x$, where $a \in K$ and $x \in Q$.*

Proof. (i) If $g \in i(K) \cap j(Q)$, then $g = i(a) = j(x)$ for $a \in K$ and $x \in Q$. Now $g = j(x)$ implies $p(g) = pj(x) = x$, while $g = i(a)$ implies $p(g) = pi(a) = 0$. Therefore, $x = 0$ and $g = j(x) = 0$.

If $g \in G$, then $p(g) = pjp(g)$ (because $pj = 1_Q$), and so $g - (jp(g)) \in \ker p = \operatorname{im} i$; hence, there is $a \in K$ with $g - (jp(g)) = i(a)$, and so $g = i(a) + j(pg) \in i(K) + j(Q)$.

(ii) Every element $g \in G$ has a factorization $g = i(a) + j(pg)$ because $G = i(K) + j(Q)$. To prove uniqueness, suppose that $i(a) + j(x) = i(b) + j(y)$, where $b \in K$ and $y \in Q$. Then $-i(b) + i(a) = j(y) - j(x) \in i(K) \cap j(Q) = \{0\}$, so that $i(a) = i(b)$ and $j(x) = j(y)$.

(iii) Necessity is the special case of (ii) when both i and j are inclusions. Conversely, each $g \in G$ has a unique factorization $g = ax$ for $a \in K$ and $x \in Q$; define $p\colon G \to Q$ by $p(ax) = x$. It is easy to check that p is a surjective homomorphism with $\ker p = K$. •

A semidirect product is so called because a direct product G of K and Q requires, in addition to $KQ = G$, and $K \cap Q = \{1\}$, that both subgroups K and Q be normal; here, only one subgroup must be normal.

Definition. If $K \le G$ and $C \le G$ satisfies $C \cap K = \{1\}$ and $KC = G$, then C is called a ***complement*** of K.

In a semidirect product G, the subgroup K is normal; on the other hand, the image $j(Q)$, which Proposition 10.5 shows to be a complement of K, may not be normal. For example, if $G = S_3$ and $K = A_3 = \langle (1\ 2\ 3) \rangle$, we may take $C = \langle \tau \rangle$, where τ is any transposition in S_3; this example also shows that complements need not be unique. However, any two complements of K are isomorphic, for any complement of K is isomorphic to G/K.

The definition of semidirect product allows the kernel K to be nonabelian, and such groups arise naturally. For example, the symmetric group S_n is a semidirect product of the alternating group A_n by $\mathbb{I}_2$. In order to keep hypotheses uniform, however, let us assume in the text (except in some exercises) that K is abelian, even though this assumption is not always needed.

Example 10.6.

(i) A direct product $K \times Q$ is a semidirect product of K by Q (and also of Q by K).

(ii) An abelian group G is a semidirect product if and only if it is a direct product (usually called a direct sum), for every subgroup of an abelian group is normal.

(iii) The dihedral group D_{2n} is a semidirect product of $\mathbb{I}_n$ by $\mathbb{I}_2$. If $D_{2n} = \langle a, b \rangle$, where $a^n = 1$, $b^2 = 1$, and $bab = a^{-1}$, then $\langle a \rangle$ is a normal subgroup having $\langle b \rangle$ as a complement.

(iv) Every Frobenius group is a semidirect product of its Frobenius kernel by its Frobenius complement.

(v) Let $G = \mathbb{H}^\times$, the multiplicative group of nonzero quaternions. It is easy to see that if $\mathbb{R}^+$ is the multiplicative group of positive reals, then the *norm* $N\colon G \to \mathbb{R}^+$, given by

$$N(a + bi + cj + dk) = a^2 + b^2 + c^2 + d^2,$$

is a homomorphism. There is a "polar decomposition" $h = rs$, where $r > 0$ and $s \in \ker N$, and G is a semidirect product of $\ker N$ by $\mathbb{R}^+$. (The normal subgroup $\ker N$ is the 3-sphere.) In Exercise 10.4, we will see that $\ker N \cong SU(2, \mathbb{C})$, the special unitary group.

(vi) Cyclic groups of prime power order are *not* semidirect products, for they cannot be a direct sum of two proper subgroups. ◂

Definition. Let K be a Q-module. An extension G of K by Q ***realizes the operators*** if, for all $x \in Q$ and $a \in K$, we have

$$xa = \ell(x) + a - \ell(x);$$

that is, the given scalar multiplication of $\mathbb{Z}Q$ on K coincides with the scalar multiplication of Corollary 10.4 arising from conjugation.

Here is the construction.

Definition. Let Q be a group and let K be a Q-module. Define

$$G = K \rtimes Q$$

to be the set of all ordered pairs $(a, x) \in K \times Q$ with the operation

$$(a, x) + (b, y) = (a + xb, xy).$$

Notice that $(a, 1) + (0, x) = (a, x)$ in $K \rtimes Q$.

Proposition 10.7. *Given a group Q and a Q-module K, then $G = K \rtimes Q$ is a semidirect product of K by Q that realizes the operators.*

Proof. We begin by proving that G is a group. For associativity,

$$\begin{aligned}[(a, x) + (b, y)] + (c, z) &= (a + xb, xy) + (c, z)\\ &= (a + xb + (xy)c, (xy)z).\end{aligned}$$

On the other hand,

$$\begin{aligned}(a, x) + [(b, y) + (c, z)] &= (a, x) + (b + yc, yz)\\ &= (a + x(b + yc), x(yz)).\end{aligned}$$

Of course, $(xy)z = x(yz)$, because of associativity in Q. The first coordinates are also equal: Since K is a Q-module, we have

$$x(b + yc) = xb + x(yc) = xb + (xy)c.$$

Thus, the operation is associative. The identity element of G is $(0, 1)$, for

$$(0, 1) + (a, x) = (0 + 1a, 1x) = (a, x),$$

and the inverse of (a, x) is $(-x^{-1}a, x^{-1})$, for

$$(-x^{-1}a, x^{-1}) + (a, x) = (-x^{-1}a + x^{-1}a, x^{-1}x) = (0, 1).$$

Therefore, G is a group, by Exercise 2.22 on page 61.

Define a function $p\colon G \to Q$ by $p\colon (a, x) \mapsto x$. Since the only "twist" occurs in the first coordinate, p is a surjective homomorphism with $\ker p = \{(a, 1)\colon a \in K\}$. If we define $i\colon K \to G$ by $i\colon a \mapsto (a, 1)$, then

$$0 \to K \xrightarrow{i} G \xrightarrow{p} Q \to 1$$

is an extension. Define $j\colon Q \to G$ by $j\colon x \mapsto (0, x)$. It is easy to see that j is a homomorphism, for $(0, x) + (0, y) = (0, xy)$. Now $pjx = p(0, x) = x$, so that $pj = 1_Q$, and the extension splits; that is, G is a semidirect product of K by Q. Finally, G realizes the operators: If $x \in Q$, then every lifting of x has the form $\ell(x) = (b, x)$ for some $b \in K$, and

$$\begin{aligned}(b, x) + (a, 1) - (b, x) &= (b + xa, x) + (-x^{-1}b, x^{-1}) \\ &= (b + xa + x(-x^{-1}b), xx^{-1}) \\ &= (b + xa - b, 1) \\ &= (xa, 1). \quad \bullet\end{aligned}$$

We return to the multiplicative notation for a moment. In the next proof, the reader will see that the operation in $K \rtimes Q$ arises from the identity

$$(ax)(by) = a(xbx^{-1})xy.$$

Theorem 10.8. *Let K be an abelian group. If a group G is a semidirect product of K by a group Q, then there is a Q-module structure on K so that $G \cong K \rtimes Q$.*

Proof. Regard G as a group with normal subgroup K that has Q as a complement. We continue writing G additively (even though it may not be abelian), and so will now write its subgroup Q additively as well. If $a \in K$ and $x \in Q$, define

$$xa = x + a - x;$$

that is, xa is the conjugate of a by x. By Proposition 10.5, each $g \in G$ has a unique expression as $g = a + x$, where $a \in K$ and $x \in Q$. It follows that $\varphi\colon G \to K \rtimes Q$, defined by $\varphi\colon a + x \mapsto (a, x)$, is a bijection. We now show that φ is an isomorphism.

$$\begin{aligned}\varphi((a + x) + (b + y)) &= \varphi(a + x + b + (-x + x) + y) \\ &= \varphi(a + (x + b - x) + x + y) \\ &= (a + xb, x + y)\end{aligned}$$

The definition of addition in $K \rtimes Q$ now gives

$$\begin{aligned}(a + xb, x + y) &= (a, x) + (b, y) \\ &= \varphi(a + x) + \varphi(b + y). \quad \bullet\end{aligned}$$

We now use semidirect products to construct some groups.

Example 10.9.
If $K = \langle a \rangle \cong \mathbb{I}_3$, then an automorphism of K is completely determined by the image of the generator a; either $a \mapsto a$ and the automorphism is 1_K, or $a \mapsto 2a$. Therefore, $\text{Aut}(K) \cong \mathbb{I}_2$; let us denote its generator by φ, so that $\varphi(a) = 2a$ and $\varphi(2a) = a$; that is, φ multiplies by 2. Let $Q = \langle x \rangle \cong \mathbb{I}_4$, and define $\theta \colon Q \to \text{Aut}(K)$ by $\theta_x = \varphi$; hence

$$xa = 2a \text{ and } x2a = a.$$

The group

$$T = \mathbb{I}_3 \rtimes \mathbb{I}_4$$

is a group of order 12. If we define $s = (2a, x^2)$ and $t = (0, x)$, then the reader may check that

$$6s = 0 \text{ and } 2t = 3s = 2(s + t).$$

The reader knows four other groups of order 12. The fundamental theorem says there are two abelian groups of this order: $\mathbb{I}_{12} \cong \mathbb{I}_3 \times \mathbb{I}_4$ and $\mathbb{I}_2 \times \mathbb{I}_6 \cong \mathbf{V} \times \mathbb{I}_3$. Two nonabelian groups of order 12 are A_4 and $S_3 \times \mathbb{I}_2$ (Exercise 10.7 on page 794 asks the reader to prove that $A_4 \not\cong S_3 \times \mathbb{I}_2$). The group T just constructed is a new example, and Exercise 10.17 on page 812 says that every group of order 12 is isomorphic to one of these five. [Note that Exercise 2.85(ii) on page 113 states that $D_{12} \cong S_3 \times \mathbb{I}_2$.] ◂

Example 10.10.
Let p be a prime and let $K = \mathbb{I}_p \oplus \mathbb{I}_p$. Hence, K is a vector space over $\mathbb{F}_p$, and so $\text{Aut}(K) \cong \text{GL}(K)$. We choose a basis a, b of K, and this gives an isomorphism $\text{Aut}(K) \cong \text{GL}(2, p)$. Let $Q = \langle x \rangle$ be a cyclic group of order p.

Define $\theta \colon Q \to \text{GL}(2, p)$ by

$$\theta \colon x^n \mapsto \begin{bmatrix} 1 & 0 \\ n & 1 \end{bmatrix}$$

for all $n \in \mathbb{Z}$. Thus,

$$xa = a + b \text{ and } xb = b.$$

It is easy to check that the commutator $x + a - x - a = xa - a = b$, and so $G = K \rtimes Q$ is a group of order p^3 with $G = \langle a, b, x \rangle$; these generators satisfy relations

$$pa = pb = px = 0, \quad b = [x, a], \quad \text{and } [b, a] = 0 = [b, x].$$

If p is odd, then we have the nonabelian group of order p^3 and exponent p in Proposition 5.45. If $p = 2$, then $|G| = 8$, and the reader is asked to prove, in Exercise 10.8 on page 794, that $G \cong D_8$; that is, $D_8 \cong \mathbf{V} \rtimes \mathbb{I}_2$. In Example 10.6(iii), we saw that D_8 is a semidirect product of $\mathbb{I}_4$ by $\mathbb{I}_2$. Thus, $\mathbf{V} \rtimes \mathbb{I}_2 \cong \mathbb{I}_4 \rtimes \mathbb{I}_2$, and so a group can have different factorizations as a semidirect product. ◂

Example 10.11.
Let k be a field and let $k^\times$ be its multiplicative group. Now $k^\times$ acts on k by multiplication (if $a \in k$ and $a \neq 0$, then the additive homomorphism $x \mapsto ax$ is an automorphism whose inverse is $x \mapsto a^{-1}x$). Therefore, the semidirect product $k \rtimes k^\times$ is defined. In particular, if $(b, a), (d, c) \in k \rtimes k^\times$, then

$$(b, a) + (d, c) = (ad + b, ac).$$

Recall that an *affine map* is a function $f : k \to k$ of the form $f : x \mapsto ax + b$, where $a, b \in k$ and $a \neq 0$, and the collection of all affine maps under composition is the group $\mathrm{Aff}(1, k)$. Note that if $g(x) = cx + d$, then

$$\begin{aligned}(f \circ g)(x) &= f(cx + d)\\ &= a(cx + d) + b\\ &= (ac)x + (ad + b).\end{aligned}$$

It is now easy to see that the function $\varphi : (b, a) \mapsto f$, where $f(x) = ax + b$, is an isomomorphism $k \rtimes k^\times \to \mathrm{Aff}(1, k)$. ◂

EXERCISES

In the first three exercises, the group K need not be abelian; in all other exercises, it is assumed to be abelian.

10.1 Kernels in this exercise may not be abelian groups.

(i) Prove that $\mathrm{SL}(2, \mathbb{F}_5)$ is an extension of $\mathbb{I}_2$ by A_5 which is not a semidirect product.

(ii) If k is a field, prove that $\mathrm{GL}(n, k)$ is a semidirect product of $\mathrm{SL}(n, k)$ by $k^\times$.

Hint. A complement consists of all matrices $\mathrm{diag}\{1, \ldots, 1, a\}$ with $a \in k^\times$.

10.2 Let G be a group of order mn, where $(m, n) = 1$. Prove that a normal subgroup K of order m has a complement in G if and only if there exists a subgroup $C \leq G$ of order n. (Kernels in this exercise may not be abelian groups.)

10.3 (***Baer***) Prove that a group G is injective[4] in the category of all groups if and only if $G = \{1\}$. (Kernels in this exercise may not be abelian groups.)
Hint. Let A be free with basis $\{x, y\}$, and let B be the semidirect product $B = A \rtimes \langle z \rangle$, where z is an element of order 2 that acts on A by $zxz = y$ and $zyz = x$.

10.4 Let $SU(2)$ be the ***special unitary group*** consisting of all complex matrices $\left[\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right]$ of determinant 1 such that

$$a\overline{b} + c\overline{d} = 0, \quad a\overline{a} + b\overline{b} = 1, \quad c\overline{c} + d\overline{d} = 1.$$

If S is the subgroup of $\mathbb{H}^\times$ in Example 10.6(v), prove that $S \cong SU(2)$.
Hint. Use Exercise 8.2 on page 531.

---

[4]The term *injective* had not yet been coined when R. Baer, who introduced the notion of injective module, proved this result. After recognizing that injective groups are duals of free groups, he jokingly called such groups ***fascist***, and he was pleased to note that they are trivial.

10.5 Give an example of a split extension of groups

$$1 \to K \xrightarrow{i} G \xrightarrow{p} Q \to 1$$

for which there does not exist a homomorphism $q: G \to K$ with $qi = 1_K$. Compare with Exercise 7.17.

10.6 Prove that $\mathbf{Q}$, the group of quaternions, is not a semidirect product.
Hint. Recall that $\mathbf{Q}$ has a unique element of order 2.

10.7 (i) Prove that $A_4 \not\cong S_3 \times \mathbb{I}_2$.
Hint. Use Proposition 2.64 saying that A_4 has no subgroup of order 6.
(ii) Prove that no two of the nonabelian groups of order 12: A_4, $S_3 \times \mathbb{I}_2$, and T are isomorphic. (See Example 10.9.)
(iii) The affine group $\mathrm{Aff}(1, \mathbb{F}_4)$ (see Example 10.11) is a nonabelian group of order 12. Is it isomorphic to A_4, $S_3 \times \mathbb{I}_2$, or $T = \mathbb{I}_3 \rtimes \mathbb{I}_4$?

10.8 Prove that the group G of order 8 constructed in Example 10.10 is isomorphic to D_8.

10.9 If K and Q are solvable groups, prove that a semidirect product of K by Q is also solvable.

10.10 Let K be an abelian group, let Q be a group, and let $\theta: Q \to \mathrm{Aut}(K)$ be a homomorphism. Prove that $K \rtimes Q \cong K \times Q$ if and only if θ is the trivial map ($\theta_x = 1_K$ for all $x \in Q$).

10.11 (i) If K is cyclic of prime order p, prove that $\mathrm{Aut}(K)$ is cyclic of order $p - 1$.
(ii) Let G be a group of order pq, where $p > q$ are primes. If $q \nmid (p - 1)$, prove that G is cyclic. Conclude, for example, that every group of order 15 is cyclic.

10.12 Let G be an additive abelian p-group, where p is prime.
(i) If $(m, p) = 1$, prove that the function $a \mapsto ma$ is an automorphism of G.
(ii) If p is an odd prime and $G = \langle g \rangle$ is a cyclic group of order p^2, prove that $\varphi: G \to G$, given by $\varphi: a \mapsto 2a$, is the unique automorphism with $\varphi(pg) = 2pg$.

10.3 General Extensions and Cohomology

We now proceed to the study of the general extension problem: Given a group Q and an abelian group K, find all (not necessarily split) extensions G of K by Q. In light of our discussion of semidirect products, that is, of split extensions, it is reasonable to refine the problem by assuming that K is a Q-module and then to seek all those extensions realizing the operators.

One way to describe a group G is to give a multiplication table for it; that is, to list all its elements $a_1, a_2, \ldots$ and all products $a_i a_j$. Indeed, this is how we constructed semidirect products: the elements are all ordered pairs (a, x) with $a \in K$ and $x \in Q$, and multiplication (really addition, because we have chosen to write G additively) is

$$(a, x) + (b, y) = (a + xb, xy).$$

O. Schreier, in 1926, solved the extension problem in this way, and we present his solution in this section. The proof is not deep; rather, it involves manipulating and organizing a long series of elementary calculations.

We must point out, however, that Schreier's solution does not allow us to determine the number of nonisomorphic middle groups G. Of course, this last question has no easy answer. If a group G has order n, then there are $n!$ different lists of its elements and hence at most $(n!)^n$ different multiplication tables for G (there are $n!$ possibilities for each of the n rows). Suppose now that H is another group of order n. The problem of determining whether or not G and H are isomorphic is essentially the problem of comparing the families of multiplication tables of each to see if there is one for G and one for H that coincide.

Our strategy is to extract enough properties of a given extension G that will suffice to reconstruct G. Thus, we may assume that K is a Q-module, that G is an extension of K by Q that realizes the operators, and that a transversal $\ell\colon Q \to G$ has been chosen. With this initial data, we see that each $g \in G$ has a unique expression of the form

$$g = a + \ell(x), \quad a \in K \quad \text{and} \quad x \in Q;$$

this follows from G being the disjoint union of the cosets $K + \ell(x)$. Furthermore, if $x, y \in Q$, then $\ell(x) + \ell(y)$ and $\ell(xy)$ are both representatives of the same coset (we do not say these representatives are the same!), and so there is an element $f(x, y) \in K$ such that

$$\ell(x) + \ell(y) = f(x, y) + \ell(xy).$$

Definition. Given a lifting $\ell\colon Q \to G$, with $\ell(1) = 0$, of an extension G of K by Q, then a ***factor set***[5] (or *cocycle*) is a function $f\colon Q \times Q \to K$ such that

$$\ell(x) + \ell(y) = f(x, y) + \ell(xy)$$

for all $x, y \in Q$.

It is natural to choose liftings with $\ell(1) = 0$, and so we have incorporated this condition into the definition of factor set; our factor sets are often called ***normalized factor sets***.

Of course, a factor set depends on the choice of lifting ℓ. When G is a split extension, then there exists a lifting that is a homomorphism; the corresponding factor set is identically 0. Therefore, we can regard a factor set as the obstruction to a lifting being a homomorphism; that is, factor sets describe how an extension differs from being a split extension.

Proposition 10.12. *Let Q be a group, K a Q-module, and $0 \to K \to G \to Q \to 1$ an extension realizing the operators. If $\ell\colon Q \to G$ is a lifting with $\ell(1) = 0$ and $f\colon Q \times Q \to K$ is the corresponding factor set, then*

(i) *for all $x, y \in Q$,*

$$f(1, y) = 0 = f(x, 1);$$

[5] If we switch to multiplicative notation, we see that a factor set occurs in the factorization $\ell(x)\ell(y) = f(x, y)\ell(xy)$.

(ii) *the **cocycle identity** holds: For all* $x, y, z \in Q$*, we have*

$$f(x, y) + f(xy, z) = xf(y, z) + f(x, yz).$$

Proof. Set $x = 1$ in the equation that defines $f(x, y)$,

$$\ell(x) + \ell(y) = f(x, y) + \ell(xy),$$

to see that $\ell(y) = f(1, y) + \ell(y)$ [since $\ell(1) = 0$, by our new assumption], and hence $f(1, y) = 0$. Setting $y = 1$ gives the other equation of (i).

The cocycle identity follows from associativity in G. For all $x, y, z \in Q$, we have

$$\begin{aligned}[\ell(x) + \ell(y)] + \ell(z) &= f(x, y) + \ell(xy) + \ell(z) \\ &= f(x, y) + f(xy, z) + \ell(xyz).\end{aligned}$$

On the other hand,

$$\begin{aligned}\ell(x) + [\ell(y) + \ell(z)] &= \ell(x) + f(y, z) + \ell(yz) \\ &= xf(y, z) + \ell(x) + \ell(yz) \\ &= xf(y, z) + f(x, yz) + \ell(xyz). \quad \bullet\end{aligned}$$

It is more interesting that the converse is true. The next result generalizes the construction of $K \rtimes Q$ in Proposition 10.7.

Theorem 10.13. *Given a group* Q *and a* Q*-module* K*, a function* $f\colon Q \times Q \to K$ *is a factor set if and only if it satisfies the cocycle identity*[6]

$$xf(y, z) - f(xy, z) + f(x, yz) - f(x, y) = 0$$

and $f(1, y) = 0 = f(x, 1)$ *for all* $x, y, z \in Q$*.*

More precisely, there is an extension G *of* K *by* Q *realizing the operators, and there is a transversal* $\ell\colon Q \to G$ *whose corresponding factor set is* f*.*

Proof. Necessity is Proposition 10.12. For the converse, define G to be the set of all ordered pairs (a, x) in $K \times Q$ equipped with the operation

$$(a, x) + (b, y) = (a + xb + f(x, y), xy).$$

(Thus, if f is identically 0, then $G = K \rtimes Q$.) The proof that G is a group is similar to the proof of Proposition 10.7. The cocycle identity is used to prove associativity:

$$\begin{aligned}\big((a, x) + (b, y)\big) + (c, z) &= (a + xb + f(x, y), xy) + (c, z) \\ &= (a + xb + f(x, y) + xyc + f(xy, z), xyz)\end{aligned}$$

[6]Written as an alternating sum, this identity is reminiscent of the formulas describing geometric cycles as described in Section 10.1.

and

$$\begin{aligned}(a,x)+\big((b,y)+(c,z)\big) &= (a,x)+(b+yc+f(y,z),yz)\\ &= (a+xb+xyc+xf(y,z)+f(x,yz),xyz).\end{aligned}$$

The cocycle identity shows that these elements are equal.

We let the reader prove that the identity is $(0,1)$ and the inverse of (a,x) is

$$-(a,x)=(-x^{-1}a-x^{-1}f(x,x^{-1}),x^{-1}).$$

Define $p\colon G\to Q$ by $p\colon(a,x)\mapsto x$. Because the only "twist" occurs in the first coordinate, it is easy to see that p is a surjective homomorphism with $\ker p=\{(a,1):a\in K\}$. If we define $i\colon K\to G$ by $i\colon a\mapsto(a,1)$, then we have an extension $0\to K\xrightarrow{i}G\xrightarrow{p}Q\to 1$.

To see that this extension realizes the operators, we must show, for every lifting ℓ, that $xa=\ell(x)+a-\ell(x)$ for all $a\in K$ and $x\in Q$. Now $\ell(x)=(b,x)$ for some $b\in K$ and

$$\begin{aligned}\ell(x)+(a,1)-\ell(x) &= (b,x)+(a,1)-(b,x)\\ &= (b+xa,x)+(-x^{-1}b-x^{-1}f(x,x^{-1}),x^{-1})\\ &= (b+xa+x[-x^{-1}b-x^{-1}f(x,x^{-1})]+f(x,x^{-1}),1)\\ &= (xa,1).\end{aligned}$$

Finally, we must show that f is the factor set determined by ℓ. Choose the lifting $\ell(x)=(0,x)$ for all $x\in Q$. The factor set F determined by ℓ is defined by

$$\begin{aligned}F(x,y) &= \ell(x)+\ell(y)-\ell(xy)\\ &= (0,x)+(0,y)-(0,xy)\\ &= (f(x,y),xy)+(-(xy)^{-1}f(xy,(xy)^{-1}),(xy)^{-1})\\ &= (f(x,y)+xy[-(xy)^{-1}f(xy,(xy)^{-1})]+f(xy,(xy)^{-1}),xy(xy)^{-1})\\ &= (f(x,y),1). \quad \bullet\end{aligned}$$

The next result shows that we have found all the extensions of a Q-module K by a group Q.

Definition. Given a group Q, a Q-module K, and a factor set f, let $G(K,Q,f)$ denote the middle group of the extension of K by Q constructed in Theorem 10.13.

Theorem 10.14. *Let Q be a group, let K be a Q-module, and let G be an extension of K by Q realizing the operators. Then there exists a factor set $f\colon Q\times Q\to K$ with*

$$G\cong G(K,Q,f).$$

Proof. Let $\ell\colon Q \to G$ be a lifting, and let $f\colon Q \times Q \to K$ be the corresponding factor set: that is, for all $x, y \in Q$, we have

$$\ell(x) + \ell(y) = f(x, y) + \ell(xy).$$

Since G is the disjoint union of the cosets, $G = \bigcup_{x \in Q} K + \ell(x)$, each $g \in G$ has a unique expression $g = a + \ell(x)$ for $a \in K$ and $x \in Q$. Uniqueness implies that the function $\varphi\colon G \to G(K, Q, f)$, given by

$$\varphi\colon g = a + \ell(x) \mapsto (a, x),$$

is a well-defined bijection. We now show that φ is an isomorphism.

$$\begin{aligned}
\varphi(a + \ell(x) + b + \ell(y)) &= \varphi(a + \ell(x) + b - \ell(x) + \ell(x) + \ell(y)) \\
&= \varphi(a + xb + \ell(x) + \ell(y)) \\
&= \varphi(a + xb + f(x, y) + \ell(xy)) \\
&= (a + xb + f(x, y), xy) \\
&= (a, x) + (b, y) \\
&= \varphi(a + \ell(x)) + \varphi(b + \ell(y)). \quad \bullet
\end{aligned}$$

Remark. For later use, note that if $a \in K$, then $\varphi(a) = \varphi(a + \ell(1)) = (a, 1)$ and, if $x \in Q$, then $\varphi(\ell(x)) = (0, x)$. This would not be so had we chosen a lifting ℓ with $\ell(1) \neq 0$. ◄

We have now described all extensions in terms of factor sets, but factor sets are determined by liftings. Any extension has many different liftings, and so our description, which depends on a choice of lifting, must have repetitions.

Lemma 10.15. *Given a group Q and a Q-module K, let G be an extension of K by Q realizing the operators. Let ℓ and ℓ' be liftings that give rise to factor sets f and f', respectively. Then there exists a function $h\colon Q \to K$ with $h(1) = 0$ and, for all $x, y \in Q$,*

$$f'(x, y) - f(x, y) = xh(y) - h(xy) + h(x).$$

Proof. For each $x \in Q$, both $\ell(x)$ and $\ell'(x)$ lie in the same coset of K in G, and so there exists an element $h(x) \in K$ with

$$\ell'(x) = h(x) + \ell(x).$$

Since $\ell(1) = 0 = \ell'(1)$, we have $h(1) = 0$. The main formula is derived as follows:

$$\begin{aligned}
\ell'(x) + \ell'(y) &= [h(x) + \ell(x)] + [h(y) + \ell(y)] \\
&= h(x) + xh(y) + \ell(x) + \ell(y),
\end{aligned}$$

because G realizes the operators. The equations continue,

$$\begin{aligned}\ell'(x) + \ell'(y) &= h(x) + xh(y) + f(x, y) + \ell(xy) \\ &= h(x) + xh(y) + f(x, y) - h(xy) + \ell'(xy).\end{aligned}$$

By definition, f' satisfies $\ell'(x) + \ell'(y) = f'(x, y) + \ell'(xy)$. Therefore,

$$f'(x, y) = h(x) + xh(y) + f(x, y) - h(xy).$$

and so

$$f'(x, y) - f(x, y) = xh(y) - h(xy) + h(x). \quad \bullet$$

Definition. Given a group Q and a Q-module K, a function $g\colon Q \times Q \to K$ is called a ***coboundary*** if there exists a function $h\colon Q \to K$ with $h(1) = 0$ such that, for all $x, y \in Q$,

$$g(x, y) = xh(y) - h(xy) + h(x).$$

The term *coboundary* arises because its formula is an alternating sum analogous to the formula for geometric boundaries that we described in Section 10.1.

We have just shown that if f and f' are factor sets of an extension G that arise from different liftings, then $f' - f$ is a coboundary.

Definition. Given a group Q and a Q-module K, define

$$Z^2(Q, K) = \{\text{all factor sets } f\colon Q \times Q \to K\}$$

and

$$B^2(Q, K) = \{\text{all coboundaries } g\colon Q \times Q \to K\}.$$

Proposition 10.16. *Given a group Q and a Q-module K, then $Z^2(Q, K)$ is an abelian group with operation pointwise addition,*

$$f + f'\colon (x, y) \mapsto f(x, y) + f'(x, y),$$

and $B^2(Q, K)$ is a subgroup of $Z^2(Q, K)$.

Proof. To see that Z^2 is a group, it suffices to prove that $f - f'$ satisfies the two identities in Proposition 10.12. This is obvious: Just subtract the equations for f and f'.

To see that B^2 is a subgroup of Z^2, we must first show that every coboundary g is a factor set; that is, that g satisfies the two identities in Proposition 10.12. This, too, is routine and is left to the reader. Next, we must show that B^2 is a nonempty subset; but the zero function, $g(x, y) = 0$ for all $x, y \in Q$, is clearly a coboundary. Finally, we show that B^2 is closed under subtraction. If $h, h'\colon Q \to K$ show that g and g' are coboundaries, that is, $g(x, y) = xh(y) - h(xy) + h(x)$ and $g'(x, y) = xh'(y) - h'(xy) + h'(x)$, then

$$(g - g')(x, y) = x(h - h')(y) - (h - h')(xy) + (h - h')(x). \quad \bullet$$

A given extension has many liftings and, hence, many factor sets, but the difference of any two of these factor sets is a coboundary. Therefore, the following quotient group suggests itself.

Definition. The ***second cohomology group*** is defined by

$$H^2(Q,K) = Z^2(Q,K)/B^2(Q,K).$$

Definition. Given a group Q and a Q-module K, two extensions G and G' of K by Q that realize the operators are called ***equivalent*** if there is a factor set f of G and a factor set f' of G' so that $f' - f$ is a coboundary.

Proposition 10.17. *Given a group Q and a Q-module K, two extensions G and G' of K by Q that realize the operators are equivalent if and only if there exists an isomorphism $\gamma\colon G \to G'$ making the following diagram commute*:

$$\begin{array}{ccccccccc} 0 & \longrightarrow & K & \xrightarrow{i} & G & \xrightarrow{p} & Q & \longrightarrow & 1 \\ & & \downarrow{\scriptstyle 1_K} & & \downarrow{\scriptstyle \gamma} & & \downarrow{\scriptstyle 1_Q} & & \\ 0 & \longrightarrow & K & \xrightarrow{i'} & G' & \xrightarrow{p'} & Q & \longrightarrow & 1 \end{array}$$

Remark. A diagram chase shows that any homomorphism γ making the diagram commute is necessarily an isomorphism. ◂

Proof. Assume that the two extensions are equivalent. We begin by setting up notation. Let $\ell\colon Q \to G$ and $\ell'\colon Q \to G'$ be liftings, and let f, f' be the corresponding factor sets; that is, for all $x, y \in Q$, we have

$$\ell(x) + \ell(y) = f(x,y) + \ell(xy),$$

with a similar equation for f' and ℓ'. Equivalence means that there is a function $h\colon Q \to K$ with $h(1) = 0$ and

$$f(x,y) - f'(x,y) = xh(y) - h(xy) + h(x)$$

for all $x, y \in Q$. Since $G = \bigcup_{x \in Q} K + \ell(x)$ is a disjoint union, each $g \in G$ has a unique expression $g = a + \ell(x)$ for $a \in K$ and $x \in Q$; similarly, each $g' \in G'$ has a unique expression $g' = a + \ell'(x)$.

This part of the proof generalizes that of Theorem 10.14. Define $\gamma\colon G \to G'$ by

$$\gamma(a + \ell(x)) = a + h(x) + \ell'(x).$$

This function makes the diagram commute. If $a \in K$, then

$$\gamma(a) = \gamma(a + \ell(1)) = a + h(1) + \ell'(1) = a;$$

furthermore,

$$p'\gamma(a + \ell(x)) = p'(a + h(x) + \ell'(x)) = x = p(a + \ell(x)).$$

Finally, γ is a homomorphism:

$$\begin{aligned}\gamma\big([a + \ell(x)] + [b + \ell(y)]\big) &= \gamma(a + xb + f(x, y) + \ell(xy)) \\ &= a + xb + f(x, y) + h(xy) + \ell'(xy),\end{aligned}$$

while

$$\begin{aligned}\gamma(a + \ell(x)) + \gamma(b + \ell(y)) &= \big(a + h(x) + \ell'(x)\big) + \big(b + h(y) + \ell'(y)\big) \\ &= a + h(x) + xb + xh(y) + f'(x, y) + \ell'(xy) \\ &= a + xb + \big(h(x) + xh(y) + f'(x, y)\big) + \ell'(xy) \\ &= a + xb + f(x, y) + h(xy) + \ell'(xy).\end{aligned}$$

We have used the given equation for $f - f'$ [remember that the terms other than $\ell'(xy)$ all lie in the abelian group K, and so they may be rearranged].

Conversely, assume that there exists an isomorphism γ making the diagram commute, so that $\gamma(a) = a$ for all $a \in K$ and

$$x = p(\ell(x)) = p'\gamma(\ell(x))$$

for all $x \in Q$. It follows that $\gamma\ell\colon Q \to G'$ is a lifting. Applying γ to the equation $\ell(x) + \ell(y) = f(x, y) + \ell(xy)$ that defines the factor set f, we see that γf is the factor set determined by the lifting $\gamma\ell$. But $\gamma f(x, y) = f(x, y)$ for all $x, y \in Q$ because $f(x, y) \in K$. Therefore, f is also a factor set of the second extension. On the other hand, if f' is any other factor set for the second extension, then Lemma 10.15 shows that $f - f' \in B^2$; that is, the extensions are equivalent. •

We say that the isomorphism γ in Proposition 10.17 ***implements*** the equivalence. The remark after Theorem 10.14 shows that the isomorphism $\gamma\colon G \to G(K, Q, f)$ implements an equivalence of extensions.

Example 10.18.

If two extensions of K by Q realizing the operators are equivalent, then their middle groups are isomorphic. However, the converse is false: We give an example of two inequivalent extensions with isomorphic middle groups. Let p be an odd prime, and consider the following diagram:

$$\begin{array}{ccccccccc} 0 & \longrightarrow & K & \xrightarrow{\;i\;} & G & \xrightarrow{\;\pi\;} & Q & \longrightarrow & 1 \\ & & \downarrow{\scriptstyle 1_K} & & \vdots & & \downarrow{\scriptstyle 1_Q} & & \\ 0 & \longrightarrow & K & \xrightarrow{\;i'\;} & G' & \xrightarrow{\;\pi'\;} & Q & \longrightarrow & 1 \end{array}$$

Define $K = \langle a \rangle$, a cyclic group of order p, $G = \langle g \rangle = G'$, a cyclic group of order p^2, and $Q = \langle x \rangle$, where $x = g + K$. In the top row, define $i(a) = pg$ and π to be the natural map;

in the bottom row define $i'(a) = 2pg$ and π' to be the natural map. Note that i' is injective because p is odd.

Suppose there is an isomorphism $\gamma\colon G \to G'$ making the diagram commute. Commutativity of the first square implies $\gamma(pa) = 2pa$, and this forces $\gamma(g) = 2g$, by Exercise 10.12(ii) on page 794; commutativity of the second square gives $g + K = 2g + K$; that is, $g \in K$. We conclude that the two extensions are not equivalent. ◄

The next theorem summarizes the calculations in this section.

Theorem 10.19 (Schreier). *Let Q be a group, let K be a Q-module, and let $e(Q, K)$ denote the family of all the equivalence classes of extensions of K by Q realizing the operators. There is a bijection*

$$\varphi\colon H^2(Q, K) \to e(Q, K)$$

that takes 0 *to the class of the split extension.*

Proof. Denote the equivalence class of an extension

$$0 \to K \to G \to Q \to 1$$

by $[G]$. Define $\varphi\colon H^2(Q, K) \to e(Q, K)$ by

$$\varphi\colon f + B^2 \mapsto [G(K, Q, f)],$$

where f is a factor set of the extension and the target extension is that constructed in Theorem 10.13.

First, φ is a well-defined injection: f and g are factor sets with $f + B^2 = g + B^2$ if and only if $[G(K, Q, f)] = [G(K, Q, g)]$, by Proposition 10.17. To see that φ is a surjection, let $[G] \in e(Q, K)$. By Theorem 10.14 and the remark following it, $[G] = [G(K, Q, f)]$ for some factor set f, and so $[G] = \varphi(f + B^2)$. Finally, the zero factor set corresponds to the semidirect product. •

If H is a group and if there is a bijection $\varphi\colon H \to X$, where X is a set, then there is a unique operation defined on X making X a group and φ an isomorphism: Given $x, y \in X$, there are $g, h \in H$ with $x = \varphi(g)$ and $y = \varphi(h)$, and we define $xy = \varphi(gh)$. In particular, there is a way to add two equivalence classes of extensions; it is called ***Baer sum*** (see Section 10.6).

Corollary 10.20. *If Q is a group, K is a Q-module, and $H^2(Q, K) = \{0\}$, then every extension of K by Q realizing the operators is a semidirect product.*

Proof. By the theorem, $e(Q, K)$ has only one element; since the split extension always exists, this one element must be the equivalence class of the split extension. Therefore, every extension of K by Q realizing the operators is split, and so its middle group is a semidirect product. •

We now apply Schreier's theorem.

Theorem 10.21. *Let G be a finite group of order mn, where $(m, n) = 1$. If K is an abelian normal subgroup of order m, then K has a complement and G is a semidirect product.*

Proof. Define $Q = G/K$. By Corollary 10.20, it suffices to prove that every factor set $f : Q \times Q \to K$ is a coboundary. Define $\sigma : Q \to K$ by

$$\sigma(x) = \sum_{y \in Q} f(x, y);$$

σ is well-defined because Q is finite and K is abelian. Now sum the cocycle identity

$$xf(y, z) - f(xy, z) + f(x, yz) - f(x, y) = 0$$

over all $z \in Q$ to obtain

$$x\sigma(y) - \sigma(xy) + \sigma(x) = nf(x, y)$$

(as z varies over all of Q, so does yz). Since $(m, n) = 1$, there are integers s and t with $sm + tn = 1$. Define $h : Q \to K$ by

$$h(x) = t\sigma(x).$$

Note that $h(1) = 0$ and

$$xh(y) - h(xy) + h(x) = f(x, y) - msf(x, y).$$

But $sf(x, y) \in K$, and so $msf(x, y) = 0$. Therefore, f is a coboundary. •

Remark. P. Hall proved that if G is a finite solvable group of order mn, where $(m, n) = 1$, then G has a subgroup of order m and any two such are conjugate. In particular, in a solvable group, every (not necessarily normal) Sylow subgroup has a complement. Because of this theorem, a (not necessarily normal) subgroup H of a finite group G is called a ***Hall subgroup*** if $(|H|, [G : H]) = 1$. Thus, Theorem 10.21 is often stated as every normal Hall subgroup of an arbitrary finite group has a complement. ◄

We now use some group theory to remove the hypothesis that K be abelian.

Theorem 10.22 (Schur-Zassenhaus[7] Lemma). *Let G be a finite group of order mn, where $(m, n) = 1$. If K is a normal subgroup of order m, then K has a complement and G is a semidirect product.*

[7] I. Schur proved this theorem, in 1904, for the special case Q cyclic. H. Zassenhaus, in 1938, proved the theorem for arbitrary finite Q.

Proof. By Exercise 10.2 on page 793, it suffices to prove that G contains a subgroup of order n; we prove the existence of such a subgroup by induction on $m \geq 1$. Of course, the base step $m = 1$ is true.

Suppose that there is a proper subgroup T of K with $\{1\} < T \lhd G$. Then $K/T \lhd G/T$ and $(G/T)/(K/T) \cong G/K$ has order n. Since $T < K$, we have $|K/T| < |K| = m$, and so the inductive hypothesis provides a subgroup $N/T \leq G/T$ with $|N/T| = n$. Now $|N| = n|T|$, where $(|T|, n) = 1$ [because $|T|$ is a divisor of $|K| = m$], so that T is a normal subgroup of N whose order and index are relatively prime. Since $|T| < |K| = m$, the inductive hypothesis provides a subgroup C of N (which is obviously a subgroup of G) of order n.

We may now assume that K is a minimal normal subgroup of G; that is, there is no normal subgroup T of G with $\{1\} < T < K$. Let p be a prime divisor of $|K|$ and let P be a Sylow p-subgroup of K. By the Frattini argument, Exercise 5.21 on page 277, we have $G = KN_G(P)$. Therefore,

$$\begin{aligned} G/K &= KN_G(P)/K \\ &\cong N_G(P)/(K \cap N_G(P)) \\ &= N_G(P)/N_K(P). \end{aligned}$$

Hence, $|N_K(P)|n = |N_K(P)||G/K| = |N_G(P)|$. If $N_G(P)$ is a proper subgroup of G, then $|N_K(P)| < m$, and induction provides a subgroup of $N_G(P) \leq G$ of order n. Therefore, we may assume that $N_G(P) = G$; that is, $P \lhd G$.

Since $\{1\} < P \leq K$ and P is normal in G, we must have $P = K$, because K is a minimal normal subgroup. But P is a p-group, and so its center, $Z(P)$, is nontrivial. By Exercise 5.19(v) on page 277, we have $Z(P) \lhd G$, and so $Z(P) = P$, again because $P = K$ is a minimal normal subgroup of G. It follows that P is abelian, and we have reduced the problem to Theorem 10.21. •

Corollary 10.23. *If a finite group G has a normal Sylow p-subgroup P, for some prime divisor p of $|G|$, then G is a semidirect product; more precisely, P has a complement.*

Proof. The order and index of a Sylow subgroup are relatively prime. •

There is another part of the Schur-Zassenhaus lemma that we have not stated: If K is a normal subgroup of G whose order and index are relatively prime, then any two complements of K are conjugate subgroups. We are now going to see that there is an analog of $H^2(K, Q)$ whose vanishing implies conjugacy of complements when K is abelian. This group, $H^1(K, Q)$, arises, as did $H^2(K, Q)$, from a series of elementary calculations.

We begin with a computational lemma. Let Q be a group, let K be a Q-module, and let $0 \to K \to G \to Q \to 1$ be a split extension. Choose a lifting $\ell\colon Q \to G$, so that every element $g \in G$ has a unique expression of the form

$$g = a + \ell x.$$

where $a \in K$ and $x \in Q$.

Definition. An automorphism φ of a group G ***stabilizes*** an extension $0 \to K \to G \to Q \to 1$ if the following diagram commutes:

$$\begin{array}{ccccccccc} 0 & \longrightarrow & K & \xrightarrow{i} & G & \xrightarrow{p} & Q & \longrightarrow & 1 \\ & & \downarrow{\scriptstyle 1_K} & & \downarrow{\scriptstyle \varphi} & & \downarrow{\scriptstyle 1_Q} & & \\ 0 & \longrightarrow & K & \xrightarrow{i} & G & \xrightarrow{p} & Q & \longrightarrow & 1 \end{array}$$

The set of all stabilizing automorphisms of an extension of K by Q, where K is a Q-module, form a group under composition, denoted by

$$\mathrm{Stab}(Q, K).$$

Note that a stabilizing automorphism is an isomorphism that implements an equivalence of an extension with itself. We shall see, in Proposition 10.26, that $\mathrm{Stab}(Q, K)$ does not depend on the extension.

Proposition 10.24. *Let Q be a group, let K be a Q-module, and let*

$$0 \to K \xrightarrow{i} G \xrightarrow{p} Q \to 1$$

be a split extension. If $\ell\colon Q \to G$ is a lifting, then every stabilizing automorphism $\varphi\colon G \to G$ has the form

$$\varphi(a + \ell x) = a + d(x) + \ell x,$$

where $d(x) \in K$ is independent of the choice of lifting ℓ. Moreover, this formula defines a stabilizing automorphism if and only if, for all $x, y \in Q$, the function $d\colon Q \to K$ satisfies

$$d(xy) = d(x) + xd(y).$$

Proof. If φ is stabilizing, then $\varphi i = i$, where $i\colon K \to G$, and $p\varphi = p$. Since we are assuming that i is the inclusion [which is merely a convenience to allow us to write a instead of $i(a)$], we have $\varphi(a) = a$ for all $a \in K$. To use the second constraint on φ, suppose that $\varphi(\ell x) = d(x) + \ell y$ for some $d(x) \in K$ and $y \in Q$. Then

$$\begin{aligned} x &= p(\ell x) \\ &= p\varphi(\ell x) \\ &= p(d(x) + \ell y) \\ &= y; \end{aligned}$$

that is, $x = y$. Therefore,

$$\varphi(a + \ell x) = \varphi(a) + \varphi(\ell x) = a + d(x) + \ell x.$$

To see that the formula for d holds, we first show that d is independent of the choice of lifting. Suppose that $\ell' \colon Q \to G$ is another lifting, so that $\varphi(\ell' x) = d'(x) + \ell' x$ for some $d'(x) \in K$. Now there is $k(x) \in K$ with $\ell' x = k(x) + \ell x$, for $p\ell' x = x = p\ell x$. Therefore,

$$\begin{aligned} d'(x) &= \varphi(\ell' x) - \ell' x \\ &= \varphi(k(x) + \ell x) - \ell' x \\ &= k(x) + d(x) + \ell x - \ell' x \\ &= d(x), \end{aligned}$$

because $k(x) + \ell x - \ell' x = 0$.

Since $d(x)$ is independent of the choice of lifting ℓ, and since the extension splits, we may assume that ℓ is a homomorphism: $\ell x + \ell y = \ell(xy)$. We compute $\varphi(\ell x + \ell y)$ in two ways. On the one hand,

$$\varphi(\ell x + \ell y) = \varphi(\ell(xy)) = d(xy) + \ell(xy).$$

On the other hand,

$$\begin{aligned} \varphi(\ell x + \ell y) &= \varphi(\ell x) + \varphi(\ell y) \\ &= d(x) + \ell x + d(y) + \ell y \\ &= d(x) + x d(y) + \ell(xy). \end{aligned}$$

The proof of the converse, if $\varphi(a + \ell x) = a + d(x) + \ell x$, where d satisfies the given identity, then φ is a stabilizing isomorphism, is a routine argument that is left to the reader. •

We give a name to functions like d.

Definition. Let Q be a group and let K be a Q-module. A ***derivation***[8] (or ***crossed homomorphism***) is a function $d \colon Q \to K$ such that

$$d(xy) = x d(y) + d(x).$$

The set of all derivations, $\mathrm{Der}(Q, K)$, is an abelian group under pointwise addition [if K is a trivial Q-module, then $\mathrm{Der}(Q, K) = \mathrm{Hom}(Q, K)$].

If d is a derivation, then $d(11) = 1d(1) + d(1) \in K$, and so $d(1) = 0$.

Example 10.25.
(i) If Q is a group and K is a Q-module, then a function $u \colon Q \to K$ of the form $u(x) = x a_0 - a_0$, where $a_0 \in K$, is a derivation:

$$\begin{aligned} u(x) + x u(y) &= x a_0 - a_0 + x(y a_0 - a_0) \\ &= x a_0 - a_0 + x y a_0 - x a_0 \\ &= x y a_0 - a_0 \\ &= u(xy). \end{aligned}$$

[8]Earlier, we defined a derivation of a (not necessarily associative) ring R as a function $d \colon R \to R$ with $d(xy) = d(x)y + x d(y)$. Derivations here are defined on modules, not on rings.

A derivation u of the form $u(x) = xa_0 - a_0$ is called a ***principal derivation***.

If the action of Q on K is conjugation, $xa = x + a - x$, then

$$xa_0 - a_0 = x + a_0 - x - a_0;$$

that is, $xa_0 - a_0$ is the commutator of x and a_0.

(ii) It is easy to check that the set $\text{PDer}(Q, K)$ of all the principal derivations is a subgroup of $\text{Der}(Q, K)$. ◂

Recall that $\text{Stab}(Q, K)$ denotes the group of all the stabilizing automorphisms of an extension of K by Q.

Proposition 10.26. *If Q is a group, K is a Q-module, and $0 \to K \to G \to Q \to 1$ is a split extension, then there is an isomorphism $\text{Stab}(Q, K) \to \text{Der}(Q, K)$.*

Proof. Let φ be a stabilizing automorphism. If $\ell \colon Q \to G$ is a lifting, then Proposition 10.24 says that $\varphi(a + \ell x) = a + d(x) + \ell x$, where d is a derivation. Since this proposition further states that d is independent of the choice of lifting, $\varphi \mapsto d$ is a well-defined function $\text{Stab}(Q, K) \to \text{Der}(Q, K)$, which is easily seen to be a homomorphism.

To see that this map is an isomorphism, we construct its inverse. If $d \in \text{Der}(Q, K)$, define $\varphi \colon G \to G$ by $\varphi(a + \ell x) = a + d(x) + \ell x$. Now φ is stabilizing, by Proposition 10.24, and $d \mapsto \varphi$ is the desired inverse function. •

It is not obvious from its definition that $\text{Stab}(Q, K)$ is abelian, for its binary operation is composition. However, $\text{Stab}(Q, K)$ is abelian, for $\text{Der}(Q, K)$ is.

Recall that an automorphism φ of a group G is called an *inner automorphism* if it is a conjugation; that is, there is $c \in G$ with $\varphi(g) = c + g - c$ for all $g \in G$ (if G is written additively).

Lemma 10.27. *Let $0 \to K \to G \to Q \to 1$ be a split extension, and let $\ell \colon Q \to G$ be a lifting. Then a function $\varphi \colon G \to G$ is an inner stabilizing automorphism by some $a_0 \in K$ if and only if*

$$\varphi(a + \ell x) = a + xa_0 - a_0 + \ell x.$$

Proof. If we write $d(x) = xa_0 - a_0$, then $\varphi(a + \ell x) = a + d(x) + \ell x$. But d is a (principal) derivation, and so φ is a stabilizing automorphism, by Proposition 10.24. Finally, φ is conjugation by $-a_0$, for

$$-a_0 + (a + \ell x) + a_0 = -a_0 + a + xa_0 + \ell x = \varphi(a + \ell x).$$

Conversely, assume that φ is a stabilizing conjugation. That φ is stabilizing says that $\varphi(a + \ell x) = a + d(x) + \ell x$; that φ is conjugation says that there is $b \in K$ with $\varphi(a + \ell x) = b + a + \ell x - b$. But $b + a + \ell x - b = b + a - xb + \ell x$, so that $d(x) = b - xb$, as desired. •

Definition. If Q is a group and K is a Q-module, define

$$H^1(Q, K) = \mathrm{Der}(Q, K)/\mathrm{PDer}(Q, K),$$

where $\mathrm{PDer}(Q, K)$ is the subgroup of $\mathrm{Der}(Q, K)$ consisting of all the principal derivations.

Proposition 10.28. *Let $0 \to K \to G \to Q \to 1$ be a split extension, and let C and C' be complements of K in G. If $H^1(Q, K) = \{0\}$, then C and C' are conjugate.*

Proof. Since G is a semidirect product, there are liftings $\ell\colon Q \to G$, with image C, and $\ell'\colon Q \to G$, with image C', which are homomorphisms. Thus, the factor sets f and f' determined by each of these liftings is identically zero, and so $f' - f = 0$. But Lemma 10.15 says that there exists $h\colon Q \to K$, namely, $h(x) = \ell' x - \ell x$, with

$$0 = f'(x, y) - f(x, y) = xh(y) - h(xy) + h(x);$$

thus, h is a derivation. Since $H^1(Q, K) = \{0\}$, h is a principal derivation: there is $a_0 \in K$ with

$$\ell' x - \ell x = h(x) = xa_0 - a_0$$

for all $x \in Q$. Since addition in G satisfies $\ell' x - a_0 = -xa_0 + \ell' x$, we have

$$\ell x = a_0 - xa_0 + \ell' x = a_0 + \ell' x - a_0.$$

But $\mathrm{im}\, \ell = C$ and $\mathrm{im}\, \ell' = C'$, and so C and C' are conjugate via a_0. •

We can now supplement the Schur–Zassenhaus theorem.

Theorem 10.29. *Let G be a finite group of order mn, where $(m, n) = 1$. If K is an abelian normal subgroup of order m, then G is a semidirect product of K by G/K, and any two complements of K are conjugate.*

Proof. By Proposition 10.28, it suffices to prove that $H^1(Q, K) = \{0\}$, where $Q = G/K$. Note, first, that $|Q| = |G|/|K| = mn/m = n$.

Let $d\colon Q \to K$ be a derivation: for all $x, y \in Q$, we have

$$d(xy) = xd(y) + d(x).$$

Sum this equation over all $y \in Q$ to obtain

$$\Delta = x\Delta + nd(x),$$

where $\Delta = \sum_{y \in Q} d(y)$ (as y varies over Q, so does xy). Since $(m, n) = 1$, there are integers s and t with $sn + tm = 1$. Hence,

$$d(x) = snd(x) + tmd(x) = snd(x),$$

because $d(x) \in K$ and so $md(x) = 0$. Therefore,

$$d(x) = s\Delta - xs\Delta.$$

Setting $a_0 = -s\Delta$, we see that d is a principal derivation. •

Removing the assumption in Theorem 10.29 that K is abelian is much more difficult than removing this assumption in Theorem 10.21. We first prove that complements are conjugate if either K or Q is a solvable group. Since $|Q|$ and $|K|$ are relatively prime, at least one of K or Q has odd order. The Feit–Thompson theorem, which says that every group of odd order is solvable, now completes the proof.

There are other applications of homology in group theory besides the Schur–Zassenhaus lemma. For example, if G is a group, $a \in G$, and $\gamma_a \colon g \mapsto aga^{-1}$ is conjugation by a, then $\gamma_a^n \colon g \mapsto a^n g a^{-n}$ for all n. Hence, if a has prime order p and $a \notin Z(G)$, then γ_a is an automorphism of order p. A theorem of W. Gaschütz uses cohomology to prove that every finite nonabelian p-group has an automorphism of order p that is not conjugation by an element of G.

Let us contemplate the formulas that have arisen.

$$\begin{aligned}
\text{factor set}: &\quad 0 = xf(y,z) - f(xy,z) + f(x,yz) - f(x,y)\\
\text{coboundary}: &\quad f(x,y) = xh(y) - h(xy) + h(x)\\
\text{derivation}: &\quad 0 = xd(y) - d(xy) + d(x)\\
\text{principal derivation}: &\quad d(x) = xa_0 - a_0
\end{aligned}$$

All these formulas involve alternating sums; factor sets and derivations seem to be in kernels, and coboundaries and principal derivations seem to be in images. Let us make this more precise.

Denote the cartesian product of n copies of Q by Q^n; for clarity, we denote an element of Q^n by $[x_1, \ldots, x_n]$ instead of by $(x_1, \ldots, x_n)$. Factor sets and coboundaries are certain functions $Q^2 \to K$, and derivations are certain functions $Q^1 \to K$. Let F_n be the free left $\mathbb{Z}Q$-module with basis Q^n. By the definition of basis, every function $f \colon Q^n \to K$ gives a unique Q-homomorphism $\widetilde{f} \colon F_n \to K$ extending f, for K is a Q-module; that is, if $\mathbf{Set}(Q^n, K)$ denotes the family of all functions $Q^n \to K$ in the category of sets, then $f \mapsto \widetilde{f}$ gives a bijection

$$\mathbf{Set}(Q^n, K) \to \mathrm{Hom}_{\mathbb{Z}Q}(F_n, K).$$

The inverse of this function is restriction

$$\mathrm{res} \colon \mathrm{Hom}_{\mathbb{Z}Q}(F_n, K) \to \mathbf{Set}(Q^n, K),$$

defined by $\mathrm{res} \colon g \mapsto g|Q^n$.

We now define maps that are suggested by the various formulas:

$$\begin{aligned}
d_3 \colon F_3 \to F_2: &\quad d_3[x,y,z] = x[y,z] - [xy,z] + [x,yz] - [x,y];\\
d_2 \colon F_2 \to F_1: &\quad d_2[x,y] = x[y] - [xy] + [x].
\end{aligned}$$

In fact, we define one more map: let $Q^0 = \{1\}$ be a 1-point set, so that $F_0 = \mathbb{Z}Q$ is the free Q-module on the single generator, 1. Now define

$$d_1 \colon F_1 \to F_0: \quad d_1[x] = x - 1.$$

We have defined each of d_3, d_2, and d_1 on bases of free modules, and so each extends to a Q-map.

Proposition 10.30. *The sequence*

$$F_3 \xrightarrow{d_3} F_2 \xrightarrow{d_2} F_1 \xrightarrow{d_1} F_0$$

is an exact sequence of Q-modules.

Sketch of Proof. We will only check that $d_1 d_2 = 0$ and $d_2 d_3 = 0$; that is, $\operatorname{im} d_2 \subseteq \ker d_1$ and $\operatorname{im} d_3 \subseteq \ker d_2$. The (trickier) reverse inclusions will be proved in Theorem 10.117 after we introduce some homological algebra.

$$\begin{aligned} d_1 d_2[x, y] &= d_1(x[y] - [xy] + [x]) \\ &= x d_1[y] - d_1[xy] + d_1[x]) \\ &= x(y-1) - (xy-1) + (x-1) \\ &= 0 \end{aligned}$$

(the equation $d_1 x[y] = x d_1[y]$ holds because d_1 is a Q-map). The reader should note that this is the same calculation as in Proposition 10.16.

$$\begin{aligned} d_2 d_3[x, y, z] &= d_2(x[y, z] - [xy, z] + [x, yz] - [x, y]) \\ &= x d_2[y, z] - d_2[xy, z] + d_2[x, yz] - d_2[x, y] \\ &= x(y[z] - [yz] + [y]) - (xy[z] - [xyz] + [xy]) \\ &\qquad + (x[yz] - [xyz] + [x]) - (x[y] - [xy] + [x]) \\ &= 0 \quad \bullet \end{aligned}$$

Let us recall that if X is a set and K is a module, then functions $X \to K$ are the same as homomorphisms $F \to K$, where F is the free module having basis X: Formally, the functors $\mathbf{Set}(X, \;)$ and $\operatorname{Hom}(F, \;)$, which map ${}_{\mathbb{Z}Q}\mathbf{Mod} \to \mathbf{Set}$, are naturally equivalent. Applying the contravariant functor $\operatorname{Hom}_{\mathbb{Z}Q}(\;, K)$ to the sequence in Proposition 10.30, we obtain a (not necessarily exact) sequence

$$\operatorname{Hom}(F_3, K) \xleftarrow{d_3^*} \operatorname{Hom}(F_2, K) \xleftarrow{d_2^*} \operatorname{Hom}(F_1, K) \xleftarrow{d_1^*} \operatorname{Hom}(F_0, K);$$

inserting the bijections $\operatorname{res}\colon g \mapsto g|Q^n$ gives a commutative diagram of sets:

$$\begin{array}{ccccccc} \mathbf{Set}(Q^3, K) & \longleftarrow & \mathbf{Set}(Q^2, K) & \longleftarrow & \mathbf{Set}(Q, K) & \longleftarrow & \mathbf{Set}(\{1\}, K) \\ \uparrow{\scriptstyle \text{res}} & & \uparrow{\scriptstyle \text{res}} & & \uparrow{\scriptstyle \text{res}} & & \uparrow{\scriptstyle \text{res}} \\ \operatorname{Hom}(F_3, K) & \xleftarrow{d_3^*} & \operatorname{Hom}(F_2, K) & \xleftarrow{d_2^*} & \operatorname{Hom}(F_1, K) & \xleftarrow{d_1^*} & \operatorname{Hom}(F_0, K). \end{array}$$

We regard a function $f\colon Q^n \to K$ as the restriction of the Q-map $\widetilde{f}\colon F_n \to K$ which extends it. Suppose that $f\colon Q^2 \to K$ lies in $\ker d_3^*$. Then $0 = d_3^*(f) = f d_3$. Hence, for all $x, y, z \in Q$, we have

$$\begin{aligned} 0 &= f d_3[x, y, z] \\ &= f(x[y, z] - [xy, z] + [x, yz] - [x, y]) \\ &= x f[y, z] - f[xy, z] + f[x, yz] - f[x, y]; \end{aligned}$$

the equation $f(x[y,z]) = xf[y,z]$ holds because f is the restriction of a Q-map. Thus, f is a factor set. If f lies in $\operatorname{im} d_2^*$, then there is some $h\colon Q \to K$ with $f = d_2^*(h) = hd_2$. Thus,

$$\begin{aligned} f[x,y] &= hd_2[x,y] \\ &= h(x[y] - [xy] + [x]) \\ &= xh[y] - h[xy] + h[x]; \end{aligned}$$

the equation $h(x[y]) = xh[y]$ holds because h is the restriction of a Q-map. Thus, f is a coboundary.

A similar analysis shows that if $g\colon Q \to K$ lies in $\ker d_2^*$, then g is a derivation. Let us now compute $\operatorname{im} d_1^*$. If $k\colon \{1\} \to K$, then

$$d_1^*(k) = kd_1(x) = k((x-1)1) = (x-1)k(1),$$

because k is the restriction of a Q-map. Now $k(1)$ is merely an element of K; indeed, if we identify k with its (1-point) image $k(1) = a_0$, then we see that $d_1^*(k)$ is a principal derivation.

Observe that $d_2 d_3 = 0$ implies $d_3^* d_2^* = 0$, which is equivalent to $\operatorname{im} d_2^* \subseteq \ker d_3^*$; that is, every coboundary is a factor set, which is Proposition 10.16. Similarly, $d_1 d_2 = 0$ implies $\operatorname{im} d_1^* \subseteq \ker d_2^*$; that is, every principal derivation is a derivation, which is Example 10.25(i).

As long as we are computing kernels and images, what is $\ker d_1^*$? If $k\colon \{1\} \to K$ and $k(1) = a_0$, then $k \in \ker d_1^*$ says

$$0 = d_1^*(k) = kd_1(x) = (x-1)k(1) = (x-1)a_0,$$

so that $xa_0 = a_0$ for all $x \in Q$. We have been led to the following definition.

Definition. If Q is a group and K is a Q-module, then the submodule of ***fixed points*** is defined by

$$H^0(Q,K) = \{a \in K : xa = a \text{ for all } x \in Q\}.$$

The groups $H^2(Q,K)$, $H^1(Q,K)$, and $H^0(Q,K)$ were obtained by applying the functor $\operatorname{Hom}(\ ,K)$ to the exact sequence $F_3 \to F_2 \to F_1 \to F_0$. In algebraic topology, we would also apply the functor $\ \otimes_{\mathbb{Z}Q} K$, obtaining ***homology groups*** [the tensor product is defined because we may view the free Q-modules F_n as right Q-modules, as in Example 8.79(v)]:

$$\begin{aligned} H_0(Q,K) &= \ker(d_0 \otimes 1)/\operatorname{im}(d_1 \otimes 1); \\ H_1(Q,K) &= \ker(d_1 \otimes 1)/\operatorname{im}(d_2 \otimes 1); \\ H_2(Q,K) &= \ker(d_2 \otimes 1)/\operatorname{im}(d_3 \otimes 1). \end{aligned}$$

We can show that $H_0(Q,K)$ is the maximal Q-trivial quotient of K. In the special case $K = \mathbb{Z}$ viewed as a trivial Q-module, we see that $H_1(Q,\mathbb{Z}) \cong Q/Q'$, where Q' is the commutator subgroup of Q.

We discuss homological algebra in the next section, for it is the proper context in which to understand these constructions.

EXERCISES

10.13 Let Q be a group and let K be a Q-module. Prove that any two split extensions of K by Q realizing the operators are equivalent.

10.14 Let Q be a group and let K be a Q-module.

(i) If K and Q are finite groups, prove that $H^2(Q, K)$ is also finite.

(ii) Let $\tau(K, Q)$ denote the number of nonisomorphic middle groups G that occur in extensions of K by Q realizing the operators. Prove that

$$\tau(K, Q) \leq |H^2(Q, K)|.$$

(iii) Give an example showing that the inequality in part (ii) can be strict.

Hint. Observe that $\tau(\mathbb{I}_p, \mathbb{I}_p) = 2$ (note that the kernel is the trivial module because every group of order p^2 is abelian).

10.15 Recall Example 5.79 on page 307: a ***generalized quaternion group*** $\mathbf{Q}_n$ is a group of order 2^n, where $n \geq 3$, which is generated by two elements a and b such that

$$a^{2^{n-1}} = 1, \quad bab^{-1} = a^{-1}, \text{ and } \quad b^2 = a^{2^{n-2}}.$$

(i) Prove that $\mathbf{Q}_n$ has a unique element z of order 2 and that $Z(\mathbf{Q}_n) = \langle z \rangle$. Conclude that $\mathbf{Q}_n$ is not a semidirect product.

(ii) Prove that $\mathbf{Q}_n$ is a ***central extension*** (i.e., θ is trivial) of $\mathbb{I}_2$ by $D_{2^{n-1}}$.

(iii) Using factor sets, give another proof of the existence of $\mathbf{Q}_n$.

10.16 If p is an odd prime, prove that every group G of order $2p$ is a semidirect product of $\mathbb{I}_p$ by $\mathbb{I}_2$, and conclude that either G is cyclic or $G \cong D_{2p}$.

10.17 Show that every group G of order 12 is isomorphic to one of the following five groups:

$$\mathbb{I}_{12}, \quad \mathbf{V} \times \mathbb{I}_3, \quad A_4, \quad S_3 \times \mathbb{I}_2, \quad T,$$

where T is the group in Example 10.9.

10.18 If Q is a group and K is a Q-module, let E be a semidirect product of K by Q and let $\ell \colon G \to E$ be a lifting. Prove that $\ell(x) = (d(x), x)$, where $d \colon Q \to K$, and ℓ is a homomorphism if and only if d is a derivation.

10.19 If $U \colon {}_{\mathbb{Z}Q}\mathbf{Mod} \to \mathbf{Sets}$ is the forgetful functor (which assigns to each module its set of elements), prove that the ordered pair (Φ, U) is an adjoint pair of functors. [By Exercise 7.39(ii) on page 471, there exists a ***free functor*** $\Phi \colon \mathbf{Set} \to {}_{\mathbb{Z}Q}\mathbf{Mod}$ that assigns to each set X the free Q-module $\Phi(X)$ with basis X.]

10.20 Prove that the functors $\mathbf{Set}(X, \ \ )$ and $\mathrm{Hom}(\Phi, \ \ )$, which map ${}_{\mathbb{Z}Q}\mathbf{Mod} \to \mathbf{Set}$, are naturally equivalent, where Φ is the free functor defined in Exercise 10.19.

10.4 HOMOLOGY FUNCTORS

Let R be a ring. In this section, the word *module* will always mean "left R-module." Given a module M, there is a free module F_0 and a surjection $\varepsilon\colon F_0 \to M$; thus, there is an exact sequence

$$0 \to \Omega_1 \xrightarrow{i} F_0 \xrightarrow{\varepsilon} M \to 0,$$

where $\Omega_1 = \ker\varepsilon$ and $i\colon \Omega_1 \to F_0$ is the inclusion. This is just another way of describing a presentation of M; that is, a description of M by generators and relations. Thus, if X is a basis of F_0, then we say that X [or $\varepsilon(X)$] are generators of M and that Ω_1 are relations. The idea now is to take generators and relations of Ω_1, getting "second-order" relations Ω_2, and to iterate this construction giving a *free resolution* of M, which should be regarded as a more detailed presentation of M by generators and relations. In algebraic topology, a topological space X is replaced by a sequence of chain groups, and this sequence yields the homology groups $H_n(X)$. We are now going to replace an R-module M by a resolution of it.

Definition. A ***projective resolution*** of a module M is an exact sequence,

$$\cdots \to P_n \to P_{n-1} \to \cdots \to P_1 \to P_0 \to M \to 0,$$

in which each module P_n is projective. A ***free resolution*** is a projective resolution in which each module P_n is free.

Proposition 10.30 displayed an exact sequence of free left $\mathbb{Z}Q$-modules

$$F_3 \xrightarrow{d_3} F_2 \xrightarrow{d_2} F_1 \xrightarrow{d_1} F_0,$$

where F_n is the free Q-module with basis Q^n. The module $F_0 = \mathbb{Z}Q$ is free on the generator 1, and the map $d_1\colon F_1 \to \mathbb{Z}Q$ is given by

$$d_1\colon [x] \mapsto x - 1.$$

Proposition 10.31. *For any group Q, there is an isomorphism $\mathbb{Z}Q/\operatorname{im} d_1 \cong \mathbb{Z}$, where $\mathbb{Z}$ is regarded as a trivial Q-module.*

Proof. Define $\varepsilon\colon \mathbb{Z}Q \to \mathbb{Z}$ by

$$\varepsilon\colon \sum_{x\in Q} m_x x \mapsto \sum_{x\in Q} m_x.$$

Now ε is a Q-map, for if $x \in Q$, then $\varepsilon(x) = 1$; on the other hand, $\varepsilon(x) = \varepsilon(x \cdot 1) = x\varepsilon(1) = 1$, because $\mathbb{Z}$ is a trivial Q-module. It is clear that ε is a surjection and that $\operatorname{im} d_1 \le \ker\varepsilon$ [because $\varepsilon(x-1) = 0$]. For the reverse inclusion, if $\sum_{x\in Q} m_x x \in \ker\varepsilon$, then $\sum_{x\in Q} m_x = 0$. Hence,

$$\sum_{x\in Q} m_x x = \sum_{x\in Q} m_x x - \Big(\sum_{x\in Q} m_x\Big)1 = \sum_{x\in Q} m_x(x-1) \in \operatorname{im} d_1.$$

Therefore, $\operatorname{coker} d_1 = \mathbb{Z}Q/\operatorname{im} d_1 \cong \mathbb{Z}$. •

Thus, the exact sequence in Proposition 10.30 can be lengthened so that it ends with $\operatorname{coker} d_1 = \mathbb{Z}Q/\operatorname{im} d_1$, and so it looks like the beginning of a free resolution of the trivial Q-module $\mathbb{Z}$.

Proposition 10.32. *Every module M has a free resolution* (*and hence it has a projective resolution*).

Proof. As in Section 10.1, there is a free module F_0 and an exact sequence

$$0 \to \Omega_1 \xrightarrow{i_1} F_0 \xrightarrow{\varepsilon} M \to 0.$$

Similarly, there is a free module F_1, a surjection $\varepsilon_1 \colon F_1 \to \Omega_1$, and an exact sequence

$$0 \to \Omega_2 \xrightarrow{i_2} F_1 \xrightarrow{\varepsilon_1} \Omega_1 \to 0.$$

Define $d_1 \colon F_1 \to F_0$ to be the composite $i_1 \varepsilon_1$. It is plain that $\operatorname{im} d_1 = \Omega_1 = \ker \varepsilon$ and $\ker d_1 = \Omega_2$, so there is an exact sequence

$$\begin{array}{ccccccccc} & & & F_1 & \xrightarrow{\quad d_1 \quad} & & F_0 & \xrightarrow{\varepsilon} & M \longrightarrow 0 \\ & & \nearrow & & \searrow^{\varepsilon_1} & \nearrow_{i_1} & & & \\ 0 \longrightarrow \Omega_2 & & & & \Omega_1 & & & & \end{array}$$

Plainly, this construction can be iterated for all $n \geq 0$ (so that the ultimate exact sequence is infinitely long). •

There is a dual construction.

Definition. An ***injective resolution*** of a module M is an exact sequence,

$$0 \to M \to E^0 \to E^1 \to \cdots \to E^n \to E^{n+1} \to \cdots,$$

in which each module E^n is injective.

Proposition 10.33. *Every module M has an injective resolution.*

Proof. We use Theorem 8.104, which states that every module can be imbedded as a submodule of an injective module. Thus, there is an injective module E^0, an injection $\eta \colon M \to E^0$, and an exact sequence

$$0 \to M \xrightarrow{\eta} E^0 \xrightarrow{p} \Sigma^1 \to 0,$$

where $\Sigma^1 = \operatorname{coker} \eta$ and p is the natural map. Now repeat: there is an injective module E^1, an imbedding $\eta^1 \colon \Sigma^1 \to E^1$, yielding an exact sequence

$$\begin{array}{cccccccc} 0 \longrightarrow M \xrightarrow{\eta} E^0 & \xrightarrow{\quad d^0 \quad} & & E^1 & & & \\ & \searrow_{p} & \nearrow^{\eta^1} & & \searrow & & \\ & & \Sigma^1 & & & \Sigma^2 \longrightarrow 0 & \end{array}$$

where d^0 is the composite $d^0 = \eta^1 p$. This construction can be iterated. •

We are now going to generalize both of these definitions.

Definition. A ***complex***[9] $(\mathbf{C}_\bullet, d_\bullet)$ is a sequence of modules and maps, for every $n \in \mathbb{Z}$,

$$\mathbf{C}_\bullet = \cdots \to C_{n+1} \xrightarrow{d_{n+1}} C_n \xrightarrow{d_n} C_{n-1} \to \cdots,$$

in which $d_n d_{n+1} = 0$ for all n. The maps d_n are called ***differentiations***.

Usually, we will shorten the notation $(\mathbf{C}_\bullet, d_\bullet)$ to $\mathbf{C}_\bullet$.

Note that the equation $d_n d_{n+1} = 0$ is equivalent to

$$\operatorname{im} d_{n+1} \subseteq \ker d_n.$$

Example 10.34.

(i) Every exact sequence is a complex, for the required inclusions, $\operatorname{im} d_{n+1} \subseteq \ker d_n$, are now equalities, $\operatorname{im} d_{n+1} = \ker d_n$.

(ii) The sequence of chain groups of a triangulated space X,

$$\cdots \to C_3(X) \xrightarrow{\partial_3} C_2(X) \xrightarrow{\partial_2} C_1(X) \xrightarrow{\partial_1} C_0(X),$$

is a complex. However, a complex is supposed to have a module for every $n \in \mathbb{Z}$. We force this to be a complex by defining $C_n(X) = \{0\}$ for all negative n; there is no problem defining differentiations $d_n \colon C_n(X) \to C_{n-1}(X)$ for $n \le 0$, for there is only the zero map from any module into $\{0\}$.

(iii) In Chapter 9, we considered the de Rham complex of a connected open subset X of $\mathbb{R}^n$:

$$0 \to \Omega^0(X) \xrightarrow{d^0} \Omega^1(X) \xrightarrow{d^1} \Omega^2(X) \to \cdots \to \Omega^{n-1}(X) \xrightarrow{d^{n-1}} \Omega^n(X) \to 0,$$

where the maps are the exterior derivatives.

(iv) The ***zero complex*** $\mathbf{0}_\bullet$ is the complex $(\mathbf{C}_\bullet, d_\bullet)$ each of whose terms $C_n = \{0\}$ and, necessarily, each of whose differentiations $d_n = 0$.

(v) If $\{M_n : n \in \mathbb{Z}\}$ is any sequence of modules, then $(\mathbf{M}_\bullet, d_\bullet)$ is a complex with nth term M_n if we define $d_n = 0$ for all n.

(vi) Every homomorphism is a differentiation. If $f : A \to B$ is a homomorphism, define a complex $(\mathbf{C}_\bullet, d_\bullet)$ with $C_1 = A$, $C_0 = B$, $d_1 = f$, and all other terms and differentiations zero.

(vii) Every projective resolution of a module M,

$$\cdots \to P_1 \to P_0 \to M \to 0,$$

is a complex if we add $\{0\}$'s to the right.

[9] These are also called ***chain complexes*** in the literature.

(viii) Every injective resolution of a module M,

$$0 \to M \to E^0 \to E^1 \to \cdots,$$

is a complex if we add $\{0\}$'s to the left.

We have used a convenient notation. According to the definition of complex, differentiations lower the index: $d_n \colon C_n \to C_{n-1}$. The simplest way to satisfy the definition is to use negative indices: define $C_{-n} = E^n$, and

$$0 \to M \to C_0 \to C_{-1} \to C_{-2} \to \cdots$$

is a complex.

(ix) If $\mathbf{C}_\bullet$ is a complex,

$$\mathbf{C}_\bullet = \cdots \to C_n \xrightarrow{d_n} C_{n-1} \to \cdots,$$

and if F is an additive (covariant) functor, say, $F \colon {}_R\mathbf{Mod} \to \mathbf{Ab}$, then $F(\mathbf{C}_\bullet)$, defined by

$$F(\mathbf{C}_\bullet) = \cdots \to F(C_n) \xrightarrow{Fd_n} F(C_{n-1}) \to \cdots,$$

is also a complex:

$$0 = F(0) = F(d_n d_{n+1}) = F(d_n)F(d_{n+1});$$

the equation $0 = F(0)$ holds because F is an additive functor. Note that even if the original complex is exact, the functored complex $F(\mathbf{C}_\bullet)$ may not be exact.

(x) If F is a contravariant additive functor, it is also true that $F(\mathbf{C}_\bullet)$ is a complex, but we have to arrange notation so that differentiations lower indices by 1. In more detail, after applying F, we have

$$F(\mathbf{C}_\bullet) = \cdots \leftarrow F(C_n) \xleftarrow{Fd_n} F(C_{n-1}) \leftarrow \cdots;$$

the differentiations Fd_n increase indices by 1. Introducing negative indices almost solves the problem. If we define $X_{-n} = F(C_n)$, then the sequence is rewritten as

$$F(\mathbf{C}_\bullet) = \cdots \to X_{-n+1} \xrightarrow{Fd_n} X_{-n} \to \cdots.$$

However, the index on the map should be $-n+1$, and not n. Define

$$\delta_{-n+1} = Fd_n.$$

The relabeled sequence now reads properly:

$$F(\mathbf{C}_\bullet) = \cdots \to X_{-n+1} \xrightarrow{\delta_{-n+1}} X_{-n} \to \cdots.$$

Negative indices are awkward, however, and the following notation is customary: Change the sign of the index by raising it to a superscript: Write

$$\delta^n = \delta_{-n}.$$

The final version of the functored sequence now looks like this:

$$F(\mathbf{C}_\bullet) = \cdots \to X^{n-1} \xrightarrow{\delta^{n-1}} X^n \to \cdots . \quad ◀$$

It is convenient to consider the category of all complexes, and so we introduce its morphisms.

Definition. If $(\mathbf{C}_\bullet, d_\bullet)$ and $(\mathbf{C}'_\bullet, d'_\bullet)$ are complexes, then a ***chain map***

$$f = f_\bullet \colon (\mathbf{C}_\bullet, d_\bullet) \to (\mathbf{C}'_\bullet, d'_\bullet)$$

is a sequence of maps $f_n \colon C_n \to C'_n$ for all $n \in \mathbb{Z}$ making the following diagram commute:

$$\begin{array}{ccccccccc}
\cdots & \longrightarrow & C_{n+1} & \xrightarrow{d_{n+1}} & C_n & \xrightarrow{d_n} & C_{n-1} & \longrightarrow & \cdots \\
 & & \downarrow{\scriptstyle f_{n+1}} & & \downarrow{\scriptstyle f_n} & & \downarrow{\scriptstyle f_{n-1}} & & \\
\cdots & \longrightarrow & C'_{n+1} & \xrightarrow{d'_{n+1}} & C'_n & \xrightarrow{d'_n} & C'_{n-1} & \longrightarrow & \cdots
\end{array}$$

It is easy to check that the composite gf of two chain maps

$$f_\bullet \colon (\mathbf{C}_\bullet, d_\bullet) \to (\mathbf{C}'_\bullet, d'_\bullet) \quad \text{and} \quad g_\bullet \colon (\mathbf{C}'_\bullet, d'_\bullet) \to (\mathbf{C}''_\bullet, d''_\bullet)$$

is itself a chain map, where $(gf)_n = g_n f_n$. The identity chain map $1_{\mathbf{C}_\bullet}$ on $(\mathbf{C}_\bullet, d_\bullet)$ is the sequence of identity maps $1_{C_n} \colon C_n \to C_n$.

Definition. If R is a ring, then the category of all complexes of left R-modules is denoted by ${}_R\mathbf{Comp}$; if the ring R is understood, then we will omit the prescript R.

The category **Comp** is a preadditive category (that is, the Hom's are abelian groups and the distributive laws hold whenever possible) if we define

$$(f + g)_n = f_n + g_n \text{ for all } n \in \mathbb{Z}.$$

The following definitions imitate the construction of homology groups of triangulated spaces that we described in Section 10.1.

Definition. If $(\mathbf{C}_\bullet, d_\bullet)$ is a complex, define

$$\begin{aligned} \textbf{\textit{n-cycles}} &= Z_n(\mathbf{C}_\bullet) = \ker d_n; \\ \textbf{\textit{n-boundaries}} &= B_n(\mathbf{C}_\bullet) = \operatorname{im} d_{n+1}. \end{aligned}$$

Since the equation $d_n d_{n+1} = 0$ in a complex is equivalent to the condition

$$\operatorname{im} d_{n+1} \subseteq \ker d_n,$$

we have $B_n(\mathbf{C}_\bullet) \subseteq Z_n(\mathbf{C}_\bullet)$ for every complex $\mathbf{C}_\bullet$.

Definition. If $\mathbf{C}_\bullet$ is a complex and $n \in \mathbb{Z}$, its nth ***homology*** is

$$H_n(\mathbf{C}_\bullet) = Z_n(\mathbf{C}_\bullet)/B_n(\mathbf{C}_\bullet).$$

Example 10.35.
A complex is an exact sequence if and only if all its homology groups are $\{0\}$: that is, $H_n(\mathbf{C}_\bullet) = \{0\}$ for all n. Thus, the homology groups measure the deviation of a complex from being an exact sequence. An exact sequence is often called an ***acyclic complex***; *acyclic* means "no cycles"; that is, no cycles that are not boundaries. ◄

Example 10.36.
In Example 10.34(vi), we saw that every homomorphism $f : A \to B$ can be viewed as part of a complex $\mathbf{C}_\bullet$ with $C_1 = A$, $C_0 = B$, $d_1 = f$, and having $\{0\}$'s to the left and to the right. Now $d_2 = 0$ implies $\operatorname{im} d_2 = 0$, and $d_0 = 0$ implies $\ker d_0 = B$; it follows that

$$H_n(\mathbf{C}_\bullet) = \begin{cases} \ker f & \text{if } n = 1; \\ \operatorname{coker} f & \text{if } n = 0; \\ 0 & \text{otherwise.} \end{cases} \quad ◄$$

Proposition 10.37. *For each $n \in \mathbb{Z}$, homology $H_n : {}_R\mathbf{Comp} \to {}_R\mathbf{Mod}$ is an additive functor.*

Proof. We have just defined H_n on objects; it remains to define H_n on morphisms. If $f : (\mathbf{C}_\bullet, d_\bullet) \to (\mathbf{C}'_\bullet, d'_\bullet)$ is a chain map, define $H_n(f) : H_n(\mathbf{C}_\bullet) \to H_n(\mathbf{C}'_\bullet)$ by

$$H_n(f) : z_n + B_n(\mathbf{C}_\bullet) \mapsto f_n z_n + B_n(\mathbf{C}'_\bullet).$$

We must show that $f_n z_n$ is a cycle and that $H_n(f)$ is independent of the choice of cycle z_n; both of these follow from f being a chain map; that is, from commutativity of the following diagram:

$$\begin{array}{ccccc} C_{n+1} & \xrightarrow{d_{n+1}} & C_n & \xrightarrow{d_n} & C_{n-1} \\ \downarrow{\scriptstyle f_{n+1}} & & \downarrow{\scriptstyle f_n} & & \downarrow{\scriptstyle f_{n-1}} \\ C'_{n+1} & \xrightarrow[d'_{n+1}]{} & C'_n & \xrightarrow[d'_n]{} & C'_{n-1} \end{array}$$

First, let z be an n-cycle in $Z_n(\mathbf{C}_\bullet)$, so that $d_n z = 0$. Then commutativity of the diagram gives

$$d'_n f_n z = f_{n-1} d_n z = 0.$$

Therefore, $f_n z$ is an n-cycle.

Next, assume that $z + B_n(\mathbf{C}_\bullet) = y + B_n(\mathbf{C}_\bullet)$; hence, $z - y \in B_n(\mathbf{C}_\bullet)$; that is,

$$z - y = d_{n+1} c$$

for some $c \in C_{n+1}$. Applying f_n gives

$$f_n z - f_n y = f_n d_{n+1} c = d'_{n+1} f_{n+1} c \in B_n(\mathbf{C}'_\bullet).$$

Thus, $f_n z + B_n(\mathbf{C}'_\bullet) = f_n y + B_n(\mathbf{C}'_\bullet)$.

Let us see that H_n is a functor. It is obvious that $H_n(1_{\mathbf{C}_\bullet})$ is the identity. If f and g are chain maps whose composite gf is defined, then for every n-cycle z, we have

$$\begin{aligned} H_n(gf)\colon z + B &\mapsto (gf)_n(z + B) \\ &= g_n f_n(z + B) \\ &= H_n(g)(f_n z + B) \\ &= H_n(g)H_n(f)(z + B). \end{aligned}$$

Finally, H_n is additive: if $g\colon (\mathbf{C}_\bullet, d_\bullet) \to (\mathbf{C}'_\bullet, d'_\bullet)$ is another chain map, then

$$\begin{aligned} H_n(f + g)\colon z + B_n(\mathbf{C}_\bullet) &\mapsto (f_n + g_n)z + B_n(\mathbf{C}'_\bullet) \\ &= f_n z + g_n z + B_n(\mathbf{C}'_\bullet) \\ &= \big(H_n(f) + H_n(g)\big)(z + B_n(\mathbf{C}'_\bullet)). \quad \bullet \end{aligned}$$

Definition. We call $H_n(f)$ the ***induced map***, and we usually denote it by f_{n*}, or even by f_*.

Proposition 10.38. *Let R and A be rings, and let $T\colon {}_R\mathbf{Mod} \to {}_A\mathbf{Mod}$ be an exact additive functor. Then T commutes with homology; that is, for every complex $(\mathbf{C}_\bullet, d_\bullet) \in {}_R\mathbf{Comp}$ and for every $n \in \mathbb{Z}$, there is an isomorphism*

$$H_n(T\mathbf{C}_\bullet, Td_\bullet) \cong TH_n(\mathbf{C}_\bullet, d_\bullet).$$

Proof. Consider the commutative diagram with exact bottom row,

$$\begin{array}{ccccccccc} & & C_{n+1} & \xrightarrow{d_{n+1}} & C_n & \xrightarrow{d_n} & C_{n-1} & & \\ & & {\scriptstyle d'_{n+1}}\downarrow & & \uparrow{\scriptstyle k} & & & & \\ 0 & \longrightarrow & \operatorname{im} d_{n+1} & \xrightarrow{j} & \ker d_n & \longrightarrow & H_n(\mathbf{C}_\bullet) & \longrightarrow & 0, \end{array}$$

where j, and k are inclusions and d'_{n+1} is just d_{n+1} with its target changed from C_n to $\operatorname{im} d_{n+1}$. Applying the exact functor T gives the commutative diagram with exact bottom row

$$\begin{array}{ccccccccc} & & TC_{n+1} & \xrightarrow{Td_{n+1}} & TC_n & \xrightarrow{Td_n} & TC_{n-1} & & \\ & & {\scriptstyle Td'_{n+1}}\downarrow & & \uparrow{\scriptstyle Tk} & & & & \\ 0 & \longrightarrow & T(\operatorname{im} d_{n+1}) & \xrightarrow{Tj} & T(\ker d_n) & \longrightarrow & TH_n(\mathbf{C}_\bullet) & \longrightarrow & 0 \end{array}$$

On the other hand, because T is exact, we have $T(\operatorname{im} d_{n+1}) = \operatorname{im} T(d_{n+1})$ and $T(\ker d_n) = \ker(T d_n)$, so that the bottom row is

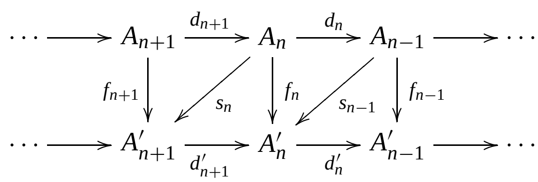
$$0 \to \operatorname{im}(T d_{n+1}) \to \ker(T d_n) \to T H_n(\mathbf{C}_\bullet) \to 0.$$

By definition, $\ker(T d_n)/\operatorname{im}(T d_{n+1}) = H_n(T\mathbf{C}_\bullet)$, and so $H_n(T\mathbf{C}_\bullet) \cong T H_n(\mathbf{C}_\bullet)$, by Proposition 8.93. •

We now introduce a notion that arises in topology.

Definition. A chain map $f: (\mathbf{C}_\bullet, d_\bullet) \to (\mathbf{C}'_\bullet, d'_\bullet)$ is ***nullhomotopic*** if, for all n, there are maps $s_n: A_n \to A'_{n+1}$ with

$$f_n = d'_{n+1} s_n + s_{n-1} d_n.$$



If $f, g: (\mathbf{C}_\bullet, d_\bullet) \to (\mathbf{C}'_\bullet, d'_\bullet)$ are chain maps, then f is ***homotopic***[10] to g, denoted by $f \simeq g$, if $f - g$ is nullhomotopic.

Proposition 10.39. *Homotopic chain maps induce the same homomorphism between homology groups: if $f, g: (\mathbf{C}_\bullet, d_\bullet) \to (\mathbf{C}'_\bullet, d'_\bullet)$ are chain maps and $f \simeq g$, then*

$$f_{*n} = g_{*n}: H_n(\mathbf{C}_\bullet) \to H_n(\mathbf{C}'_\bullet).$$

Proof. If z is an n-cycle, then $d_n z = 0$ and

$$f_n z - g_n z = d'_{n+1} s_n z + s_{n-1} d_n z = d'_{n+1} s_n z.$$

Therefore, $f_n z - g_n z \in B_n(\mathbf{C}'_\bullet)$, and so $f_{*n} = g_{*n}$. •

Definition. A complex $(\mathbf{C}_\bullet, d_\bullet)$ has a ***contracting homotopy***[11] if its identity $1_{\mathbf{C}_\bullet}$ is nullhomotopic.

---

[10] Two continuous functions $f, g: X \to Y$ are called ***homotopic*** if f can be "deformed" into g; that is, there exists a continuous $F: X \times \mathbf{I} \to Y$, where $\mathbf{I} = [0, 1]$ is the closed unit interval, with $F(x, 0) = f(x)$ and $F(x, 1) = g(x)$ for all $x \in X$. Now every continuous $f: X \to Y$ induces homomorphisms $f_*: H_n(X) \to H_n(Y)$, and one proves that if f and g are homotopic, then $f_* = g_*$. The algebraic definition of homotopy given here has been distilled from the proof of this topological theorem.

[11] A topological space is called ***contractible*** if its identity map is homotopic to a constant map.

Proposition 10.40. *A complex* $(\mathbf{C}_\bullet, d_\bullet)$ *having a contracting homotopy is acyclic*; *that is, it is an exact sequence.*

Proof. We use Example 10.35. Now $1_{\mathbf{C}_\bullet}\colon H_n(\mathbf{C}_\bullet) \to H_n(\mathbf{C}_\bullet)$ is the identity map, while $0_*\colon H_n(\mathbf{C}_\bullet) \to H_n(\mathbf{C}_\bullet)$ is the zero map. Since $1_{\mathbf{C}_\bullet} \simeq 0$, however, these maps are the same. It follows that $H_n(\mathbf{C}_\bullet) = \{0\}$ for all n; that is, $\ker d_n = \operatorname{im} d_{n+1}$ for all n, and this is the definition of exactness. •

Once we complete the free resolution of the trivial $\mathbb{Z}Q$-module $\mathbb{Z}$ whose first few terms were given in Proposition 10.30 (see also Proposition 10.31), we will prove that it is an exact sequence by showing that it has a contracting homotopy as a complex of abelian groups.

In order to study the homology functors, it is necessary to understand their domain **Comp**. Many of the constructions in ${}_R$**Mod** can also be done in the category **Comp**. We merely list the definitions and state certain properties, whose verifications are straightforward exercises for the reader.

(i) An ***isomorphism*** in **Comp** is an equivalence in this category. The reader should check that a chain map $f\colon \mathbf{C}_\bullet \to \mathbf{C}'_\bullet$ is an isomorphism if and only if $f_n\colon C_n \to C'_n$ is an isomorphism in ${}_R$**Mod** for all $n \in \mathbb{Z}$. (We must check that the sequence of inverses f_n^{-1} is a chain map; that is, that the appropriate diagram commutes.)

(ii) A complex $(\mathbf{A}_\bullet, \delta_\bullet)$ is a ***subcomplex*** of a complex $(\mathbf{C}_\bullet, d_\bullet)$ if, for every $n \in \mathbb{Z}$, we have A_n a submodule of C_n and $\delta_n = d_n | A_n$.

If $i_n\colon A_n \to C_n$ is the inclusion, then it is easy to see that $\mathbf{A}_\bullet$ is a subcomplex of $\mathbf{C}_\bullet$ if and only if $i\colon \mathbf{A}_\bullet \to \mathbf{C}_\bullet$ is a chain map.

(iii) If $\mathbf{A}_\bullet$ is a subcomplex of $\mathbf{C}_\bullet$, then the ***quotient complex*** is

$$\mathbf{C}_\bullet/\mathbf{A}_\bullet = \cdots \to C_n/A_n \xrightarrow{d''_n} C_{n-1}/A_{n-1} \to \cdots,$$

where $d''_n\colon c_n + A_n \mapsto d_n c_n + A_{n-1}$ (it must be shown that d''_n is well-defined: if $c_n + A_n = b_n + A_n$, then $d_n c_n + A_{n-1} = d_n b_n + A_{n-1}$). If $\pi_n\colon C_n \to C_n/A_n$ is the natural map, then $\pi\colon \mathbf{C}_\bullet \to \mathbf{C}_\bullet/\mathbf{A}_\bullet$ is a chain map.

(iv) If $f_\bullet\colon (\mathbf{C}_\bullet, d_\bullet) \to (\mathbf{C}'_\bullet, d'_\bullet)$ is a chain map, define

$$\mathbf{ker}\, f = \cdots \to \ker f_{n+1} \xrightarrow{\delta_{n+1}} \ker f_n \xrightarrow{\delta_n} \ker f_{n-1} \to \cdots,$$

where $\delta_n = d_n | \ker f_n$, and define

$$\mathbf{im}\, f = \cdots \to \operatorname{im} f_{n+1} \xrightarrow{\Delta_{n+1}} \operatorname{im} f_n \xrightarrow{\Delta_n} \operatorname{im} f_{n-1} \to \cdots,$$

where $\Delta_n = d'_n | \operatorname{im} f_n$. It is easy to see that $\mathbf{ker}\, f$ is a subcomplex of $\mathbf{C}_\bullet$, that $\mathbf{im}\, f$ is a subcomplex of $\mathbf{C}'_\bullet$, and that the ***first isomorphism theorem*** holds:

$$\mathbf{C}_\bullet/\mathbf{ker}\, f \cong \mathbf{im}\, f.$$

(v) A sequence of complexes and chain maps

$$\cdots \to \mathbf{C_\bullet}^{n+1} \xrightarrow{f^{n+1}} \mathbf{C_\bullet}^{n} \xrightarrow{f^{n}} \mathbf{C_\bullet}^{n-1} \to \cdots$$

is an ***exact sequence*** if, for all $n \in \mathbb{Z}$,

$$\mathbf{im}\, f^{n+1} = \mathbf{ker}\, f^{n}.$$

We may check that if $\mathbf{A_\bullet}$ is a subcomplex of $\mathbf{C_\bullet}$, then there is an exact sequence of complexes

$$\mathbf{0_\bullet} \to \mathbf{A_\bullet} \xrightarrow{i} \mathbf{C_\bullet},$$

where $\mathbf{0_\bullet}$ is the zero complex and i is the chain map of inclusions. More generally, if $i\colon \mathbf{C_\bullet} \to \mathbf{C'_\bullet}$ is a chain map, then each i_n is injective if and only if there is an exact sequence $\mathbf{0_\bullet} \to \mathbf{C_\bullet} \xrightarrow{i} \mathbf{C'_\bullet}$. Similarly, if $p\colon \mathbf{C_\bullet} \to \mathbf{C''_\bullet}$ is a chain map, then each p_n is surjective if and only if there is an exact sequence

$$\mathbf{C_\bullet} \xrightarrow{p} \mathbf{C''_\bullet} \to \mathbf{0_\bullet}.$$

The reader should realize that this notation is very compact. For example, if we write a complex as a column, then a short exact sequence of complexes is really the infinite commutative diagram with exact rows:

$$\begin{array}{ccccccccc}
 & & \downarrow & & \downarrow & & \downarrow & & \\
0 & \longrightarrow & C'_{n+1} & \xrightarrow{i_{n+1}} & C_{n+1} & \xrightarrow{p_{n+1}} & C''_{n+1} & \longrightarrow & 0 \\
 & & \downarrow{\scriptstyle d'_{n+1}} & & \downarrow{\scriptstyle d_{n+1}} & & \downarrow{\scriptstyle d''_{n+1}} & & \\
0 & \longrightarrow & C'_{n} & \xrightarrow{i_{n}} & C_{n} & \xrightarrow{p_{n}} & C''_{n} & \longrightarrow & 0 \\
 & & \downarrow{\scriptstyle d'_{n}} & & \downarrow{\scriptstyle d_{n}} & & \downarrow{\scriptstyle d''_{n}} & & \\
0 & \longrightarrow & C'_{n-1} & \xrightarrow{i_{n-1}} & C_{n-1} & \xrightarrow{p_{n-1}} & C''_{n-1} & \longrightarrow & 0 \\
 & & \downarrow & & \downarrow & & \downarrow & &
\end{array}$$

A sequence of complexes $\cdots \to \mathbf{C_\bullet}^{n+1} \xrightarrow{f^{n+1}} \mathbf{C_\bullet}^{n} \xrightarrow{f^{n}} \mathbf{C_\bullet}^{n-1} \to \cdots$ is exact if and only if

$$\cdots \to C_m^{n+1} \to C_m^{n} \to C_m^{n-1} \to \cdots$$

is an exact sequence of modules for every $m \in \mathbb{Z}$.

(vi) If $\{(\mathbf{C}^\alpha_\bullet, d^\alpha_\bullet)\}$ is a family of complexes, then their ***direct sum*** is the complex

$$\sum_\alpha \mathbf{C}^\alpha_\bullet = \cdots \to \sum_\alpha C^\alpha_{n+1} \xrightarrow{\sum_\alpha d^\alpha_n} \sum_\alpha C^\alpha_n \xrightarrow{\sum_\alpha d^\alpha_{n-1}} \sum_\alpha C^\alpha_{n-1} \to \cdots,$$

where $\sum_\alpha d^\alpha_n$ acts coordinatewise; that is, $\sum_\alpha d^\alpha_n : (c^\alpha_n) \mapsto (d^\alpha_n c^\alpha_n)$.

To summarize, we can view **Comp** as a category having virtually the same properties as the category of modules; indeed, we should view a complex as a generalized module. (Categories such as $_R$**Mod** and **Comp** are called ***abelian categories***.)

The following elementary construction is fundamental; it gives a relation between different homology modules. The proof is a series of diagram chases. Ordinarily, we would just say that the proof is routine, but, because of the importance of the result, we present (perhaps too many) details; as a sign that the proof is routine, we drop subscripts.

Proposition 10.41 (Connecting Homomorphism). *If*

$$\mathbf{0}_\bullet \to \mathbf{C}'_\bullet \xrightarrow{i} \mathbf{C}_\bullet \xrightarrow{p} \mathbf{C}''_\bullet \to \mathbf{0}_\bullet$$

is an exact sequence of complexes, then, for each $n \in \mathbb{Z}$, there is a homomorphism

$$\partial_n : H_n(\mathbf{C}''_\bullet) \to H_{n-1}(\mathbf{C}'_\bullet)$$

defined by

$$\partial_n : z''_n + B_n(\mathbf{C}''_\bullet) \mapsto i^{-1}_{n-1} d_n p^{-1}_n z''_n + B_{n-1}(\mathbf{C}'_\bullet).$$

Proof. We will make many notational abbreviations in this proof. Consider the commutative diagram having exact rows:

$$\begin{array}{ccccccccc}
 & & \downarrow & & \downarrow & & \downarrow & & \\
0 & \to & C'_{n+1} & \xrightarrow{i_{n+1}} & C_{n+1} & \xrightarrow{p_{n+1}} & C''_{n+1} & \to & 0 \\
 & & \downarrow d'_{n+1} & & \downarrow d_{n+1} & & \downarrow d''_{n+1} & & \\
0 & \to & C'_n & \xrightarrow{i_n} & C_n & \xrightarrow{p_n} & C''_n & \to & 0 \\
 & & \downarrow d'_n & & \downarrow d_n & & \downarrow d''_n & & \\
0 & \to & C'_{n-1} & \xrightarrow{i_{n-1}} & C_{n-1} & \xrightarrow{p_{n-1}} & C''_{n-1} & \to & 0 \\
 & & \downarrow & & \downarrow & & \downarrow & &
\end{array}$$

Suppose that $z'' \in C''_n$ and $d''z'' = 0$. Since p_n is surjective, there is $c \in C_n$ with $pc = z''$. Now push c down to $dc \in C_{n-1}$. By commutativity, $p_{n-1}dc = d''p_nc = d''z'' = 0$, so

that $dc \in \ker p_{n-1} = \operatorname{im} i_{n-1}$. Therefore, there is a unique $c' \in C'_{n-1}$ with $i_{n-1}c' = dc$, for i_{n-1} is an injection. Thus, $i_{n-1}^{-1} dp_n^{-1} z''$ makes sense; that is, the claim is that

$$\partial_n(z'' + B''_n) = c' + B'_{n-1}$$

is a well-defined homomorphism.

First, let us show independence of the choice of lifting. Suppose that $p_n \check{c} = z''$, where $\check{c} \in C_n$. Then $c - \check{c} \in \ker p_n = \operatorname{im} i_n$, so that there is $u' \in C'_n$ with $i_n u' = c - \check{c}$. By commutativity of the first square, we have

$$i_{n-1} d' u' = d i_n u' = dc - d\check{c}.$$

Hence, $i^{-1}dc - i^{-1}d\check{c} = d'u' \in B'_{n-1}$; that is, $i^{-1}dc + B'_{n-1} = i^{-1}d\check{c} + B'_{n-1}$. Thus, the formula gives a well-defined function

$$Z''_n \to C'_{n-1}/B'_{n-1}.$$

Second, the function $Z''_n \to C'_{n-1}/B'_{n-1}$ is a homomorphism. If $z'', z''_1 \in Z''_n$, let $pc = z''$ and $pc_1 = z''_1$. Since the definition of ∂ is independent of the choice of lifting, choose $c + c_1$ as a lifting of $z'' + z''_1$. This step may now be completed in a routine way.

Third, we show that if $i_{n-1}c' = dc$, then c' is a cycle: $0 = ddc = dic' = idc'$, and so $d'c' = 0$ because i is an injection. Hence, the formula gives a homomorphism

$$Z'' \to Z'/B' = H_{n-1}.$$

Finally, the subgroup B''_n goes into B'_{n-1}. Suppose that $z'' = d''c''$, where $c'' \in C''_{n+1}$, and let $pu = c''$, where $u \in C_{n+1}$. Commutativity gives $pdu = d''pu = d''c'' = z''$. Since $\partial(z'')$ is independent of the choice of lifting, we choose du with $pdu = z''$, and so $\partial(z'' + B'') = i^{-1}d(du) + B' = B'$. Therefore, the formula does give a homomorphism $\partial_n \colon H_n(\mathbf{C}''_\bullet) \to H_{n-1}(\mathbf{C}'_\bullet)$. •

The first question we ask is what homology functors do to a short exact sequence of complexes. The next theorem is also proved by diagram chasing and, again, we give too many details because of the importance of the result. The reader should try to prove the theorem before looking at the proof we give.

Theorem 10.42 (Long Exact Sequence). *If*

$$\mathbf{0}_\bullet \to \mathbf{C}'_\bullet \xrightarrow{i} \mathbf{C}_\bullet \xrightarrow{p} \mathbf{C}''_\bullet \to \mathbf{0}_\bullet$$

is an exact sequence of complexes, then there is an exact sequence of modules

$$\cdots \to H_{n+1}(\mathbf{C}''_\bullet) \xrightarrow{\partial_{n+1}} H_n(\mathbf{C}'_\bullet) \xrightarrow{i_*} H_n(\mathbf{C}_\bullet) \xrightarrow{p_*} H_n(\mathbf{C}''_\bullet) \xrightarrow{\partial_n} H_{n-1}(\mathbf{C}'_\bullet) \to \cdots.$$

Proof. This proof is also routine. Our notation is abbreviated, and there are six inclusions to verify.

(i) $\operatorname{im} i_* \subseteq \ker p_*$

$$p_* i_* = (pi)_* = 0_* = 0$$

(ii) $\ker p_* \subseteq \operatorname{im} i_*$

If $p_*(z + B) = pz + B'' = B''$, then $pz = d''c''$ for some $c'' \in C''_{n+1}$. But p surjective gives $c'' = pc$ for some $c \in C_{n+1}$, so that $pz = d''pc = pdc$, because p is a chain map, and so $p(z - dc) = 0$. By exactness, there is $c' \in C'_n$ with $ic' = z - dc$. Now c' is a cycle, for $id'c' = dic' = dz - ddc = 0$, because z is a cycle; since i is injective, $d'c' = 0$. Therefore,

$$i_*(c' + B') = ic' + B = z - dc + B = z + B.$$

(iii) $\operatorname{im} p_* \subseteq \ker \partial$

If $p_*(c + B) = pc + B'' \in \operatorname{im} p_*$, then $\partial(pz + B'') = z' + B'$, where $iz' = dp^{-1}pz$. Since this formula is independent of the choice of lifing of pz, let us choose $p^{-1}pz = z$. Now $dp^{-1}pz = dz = 0$, because z is a cycle. Thus, $iz' = 0$, and hence $z' = 0$, because i is injective.

(iv) $\ker \partial \subseteq \operatorname{im} p_*$

If $\partial(z'' + B'') = B'$, then $z' = i^{-1}dp^{-1}z'' \in B'$; that is, $z' = d'c'$ for some $c' \in C'$. But $iz' = id'c' = dic' = dp^{-1}z''$, so that $d(p^{-1}z'' - ic') = 0$; that is, $p^{-1}z'' - ic'$ is a cycle. Moreover, since $pi = 0$ because of exactness of the original sequence,

$$p_*(p^{-1}z'' - ic' + B) = pp^{-1}z'' - pic' + B'' = z'' + B''.$$

(v) $\operatorname{im} \partial \subseteq \ker i_*$

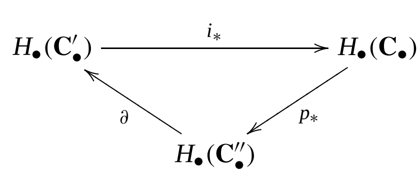
We have $i_*\partial(z'' + B'') = iz' + B'$, where $iz' = dp^{-1}z'' \in B$; that is, $i_*\partial = 0$.

(vi) $\ker i_* \subseteq \operatorname{im} \partial$

If $i_*(z' + B') = iz' + B = B$, then $iz' = dc$ for some $c \in C$. Since p is a chain map, $d''pc = pdc = piz' = 0$, by exactness of the original sequence, and so pc is a cycle. But

$$\partial(pc + B'') = i^{-1}dp^{-1}pc + B' = i^{-1}dc + B' = i^{-1}iz' + B' = z' + B'. \quad \bullet$$

Theorem 10.42 is often called the ***exact triangle*** because of the diagram



Corollary 10.43 (Snake Lemma). *Given a commutative diagram of modules with exact rows,*

$$\begin{array}{ccccccccc} 0 & \longrightarrow & A' & \longrightarrow & A & \longrightarrow & A'' & \longrightarrow & 0 \\ & & \downarrow{\scriptstyle f} & & \downarrow{\scriptstyle g} & & \downarrow{\scriptstyle h} & & \\ 0 & \longrightarrow & B' & \longrightarrow & B & \longrightarrow & B'' & \longrightarrow & 0 \end{array}$$

there is an exact sequence

$$0 \to \ker f \to \ker g \to \ker h \to \operatorname{coker} f \to \operatorname{coker} g \to \operatorname{coker} h \to 0.$$

Proof. If we view each of the vertical maps f, g, and h as a complex [as in Example 10.34(vi)], then the given commutative diagram can be viewed as a short exact sequence of complexes. The homology groups of each of these complexes has only two nonzero terms: for example, Example 10.36 shows that the homology groups of the first column are $H_1 = \ker f$, $H_0 = \operatorname{coker} f$, and all other $H_n = \{0\}$. The snake lemma now follows at once from the long exact sequence. •

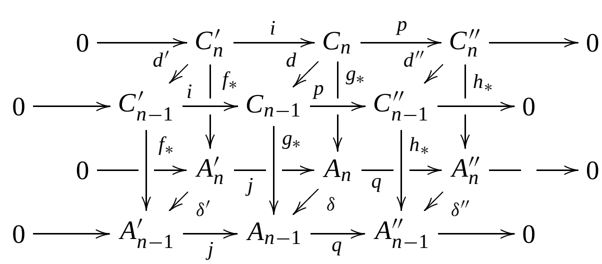
Theorem 10.44 (Naturality of ∂). *Given a commutative diagram of complexes with exact rows,*

$$\begin{array}{ccccccccc} \mathbf{0}_\bullet & \longrightarrow & \mathbf{C}'_\bullet & \xrightarrow{i} & \mathbf{C}_\bullet & \xrightarrow{p} & \mathbf{C}''_\bullet & \longrightarrow & \mathbf{0}_\bullet \\ & & \downarrow{\scriptstyle f} & & \downarrow{\scriptstyle g} & & \downarrow{\scriptstyle h} & & \\ \mathbf{0}_\bullet & \longrightarrow & \mathbf{A}'_\bullet & \xrightarrow{j} & \mathbf{A}_\bullet & \xrightarrow{q} & \mathbf{A}''_\bullet & \longrightarrow & \mathbf{0}_\bullet \end{array}$$

there is a commutative diagram of modules with exact rows,

$$\begin{array}{ccccccccccc} \cdots & \longrightarrow & H_n(\mathbf{C}'_\bullet) & \xrightarrow{i_*} & H_n(\mathbf{C}_\bullet) & \xrightarrow{p_*} & H_n(\mathbf{C}''_\bullet) & \xrightarrow{\partial} & H_{n-1}(\mathbf{C}'_\bullet) & \longrightarrow & \cdots \\ & & \downarrow{\scriptstyle f_*} & & \downarrow{\scriptstyle g_*} & & \downarrow{\scriptstyle h_*} & & \downarrow{\scriptstyle f_*} & & \\ \cdots & \longrightarrow & H_n(\mathbf{A}'_\bullet) & \xrightarrow{j_*} & H_n(\mathbf{A}_\bullet) & \xrightarrow{q_*} & H_n(\mathbf{A}''_\bullet) & \xrightarrow{\partial'} & H_{n-1}(\mathbf{A}'_\bullet) & \longrightarrow & \cdots \end{array}$$

Proof. Exactness of the rows is Theorem 10.42, while commutativity of the first two squares follows from H_n being a functor. To prove commutativity of the square involving the connecting homomorphism, let us first display the chain maps and differentiations in one (three-dimensional!) diagram:



If $z'' + B(\mathbf{C}''_\bullet) \in H_n(\mathbf{C}''_\bullet)$, we must show that

$$f_*\partial(z'' + B(\mathbf{C}''_\bullet)) = \partial' h_*(z'' + B(\mathbf{C}''_\bullet)).$$

Let $c \in C_n$ be a lifting of z''; that is, $pc = z''$. Now $\partial(z'' + B(\mathbf{C}''_\bullet)) = z' + B(\mathbf{C}'_\bullet)$, where $iz' = dc$. Hence, $f_*\partial(z'' + B(\mathbf{A}''_\bullet)) = fz' + B(\mathbf{A}'_\bullet)$. On the other hand, since h is a chain map, we have $qgc = hpc = hz''$. In computing $\partial'(hz'' + B(\mathbf{A}''_\bullet))$, we choose gc as the lifting of hz''. Hence, $\partial'(hz'' + B(\mathbf{A}''_\bullet) = u' + B(\mathbf{A}'_\bullet)$, where $ju' = \delta gc$. But

$$jfz' = giz' = gdc = \delta gc = ju',$$

and so $fz' = u'$, because j is injective. •

We shall apply these general results in the next section.

EXERCISES

10.21 If $\mathbf{C}_\bullet$ is a complex with $C_n = \{0\}$ for some n, prove that $H_n(\mathbf{C}_\bullet) = \{0\}$.

10.22 Prove that isomorphic complexes have the same homology: If $\mathbf{C}_\bullet$ and $\mathbf{D}_\bullet$ are isomorphic, then $H_n(\mathbf{C}_\bullet) \cong H_n(\mathbf{D}_\bullet)$ for all n.

10.23 Regard the map $d : \mathbb{Z} \to \mathbb{Z}$, defined by $d : m \mapsto 2m$, as a complex, as in Example 10.34(vi). Prove that it is not a projective object in the category ${}_{\mathbb{Z}}\mathbf{Comp}$ even though each of its terms is a projective $\mathbb{Z}$-module.

10.24 View $\mathbb{Z}$ as the category $\mathbf{PO}(\mathbb{Z})$ whose objects are the integers, and with exactly one morphism $n \to m$ whenever $m \le n$, with no morphisms otherwise. [If we view $\mathbb{Z}$ as a partially ordered set, then this is the associated category defined in Example 7.25(v).] Prove that a complex $(\mathbf{C}_\bullet, d_\bullet)$ is a contravariant functor $\mathbf{PO}(\mathbb{Z}) \to {}_R\,\mathbf{Mod}$, and that a chain map is a natural transformation.

10.25 In this exercise, we prove that the snake lemma implies the long exact sequence (the converse is Corollary 10.43). Consider a commutative diagram with exact rows (note that two zeros are "missing" from this diagram):

$$\begin{array}{ccccccc} & & A & \longrightarrow & B & \xrightarrow{p} & C & \longrightarrow & 0 \\ & & \downarrow{\scriptstyle\alpha} & & \downarrow{\scriptstyle\beta} & & \downarrow{\scriptstyle\gamma} & & \\ 0 & \longrightarrow & A' & \xrightarrow{i} & B' & \longrightarrow & C' & & \end{array}$$

(i) Prove that $\Delta\colon \ker\gamma \to \operatorname{coker}\alpha$, defined by

$$\Delta\colon z \mapsto i^{-1}\beta p^{-1} z + \operatorname{im}\alpha,$$

is a well-defined homomorphism.

(ii) Prove that there is an exact sequence

$$\ker\alpha \to \ker\beta \to \ker\gamma \xrightarrow{\Delta} \operatorname{coker}\alpha \to \operatorname{coker}\beta \to \operatorname{coker}\gamma.$$

(iii) Given a commutative diagram with exact rows,

$$\begin{array}{ccccccccc}
0 & \longrightarrow & A'_n & \longrightarrow & A_n & \longrightarrow & A''_n & \longrightarrow & 0 \\
 & & \Big\downarrow{\scriptstyle d'_n} & & \Big\downarrow{\scriptstyle d} & & \Big\downarrow{\scriptstyle d''_n} & & \\
0 & \longrightarrow & A'_{n-1} & \longrightarrow & A_{n-1} & \longrightarrow & A''_{n-1} & \longrightarrow & 0
\end{array}$$

prove that the following diagram is commutative and has exact rows:

$$\begin{array}{ccccccccc}
 & & A'_n/\operatorname{im} d'_{n+1} & \longrightarrow & A_n/\operatorname{im} d_{n+1} & \longrightarrow & A''_n/\operatorname{im} d''_{n+1} & \longrightarrow & 0 \\
 & & \Big\downarrow{\scriptstyle d'} & & \Big\downarrow{\scriptstyle d} & & \Big\downarrow{\scriptstyle d''} & & \\
0 & \longrightarrow & \ker d'_{n-1} & \longrightarrow & \ker d_{n-1} & \longrightarrow & \ker d''_{n-1} & &
\end{array}$$

(iv) Use part (ii) and this last diagram to give another proof of the long exact sequence.

10.26 Let $f, g\colon \mathbf{C}_\bullet \to \mathbf{C}'_\bullet$ be chain maps, and let $F\colon \mathbf{C}_\bullet \to \mathbf{C}'_\bullet$ be an additive functor. If $f \simeq g$, prove that $Ff \simeq Fg$; that is, if f and g are homotopic, then Ff and Fg are homotopic.

10.27 Let $\mathbf{0}_\bullet \to \mathbf{C}'_\bullet \xrightarrow{i} \mathbf{C}_\bullet \xrightarrow{p} \mathbf{C}''_\bullet \to \mathbf{0}_\bullet$ be an exact sequence of complexes in which $\mathbf{C}'_\bullet$ and $\mathbf{C}''_\bullet$ are acyclic; prove that $\mathbf{C}_\bullet$ is also acyclic.

10.28 Let $(\mathbf{C}_\bullet, d_\bullet)$ be a complex each of whose differentiations d_n is the zero map. Prove that $H_n(\mathbf{C}_\bullet) \cong C_n$ for all n.

10.29 (**3 × 3 *Lemma***) Given a commutative diagram in which the columns and the bottom two rows are exact sequences,

$$\begin{array}{ccccccccc}
 & & 0 & & 0 & & 0 & & \\
 & & \downarrow & & \downarrow & & \downarrow & & \\
0 & \longrightarrow & K' & \longrightarrow & K & \longrightarrow & K'' & \longrightarrow & 0 \\
 & & \downarrow & & \downarrow & & \downarrow & & \\
0 & \longrightarrow & P' & \longrightarrow & P & \longrightarrow & P'' & \longrightarrow & 0 \\
 & & \downarrow & & \downarrow & & \downarrow & & \\
0 & \longrightarrow & A' & \longrightarrow & A & \longrightarrow & A'' & \longrightarrow & 0 \\
 & & \downarrow & & \downarrow & & \downarrow & & \\
 & & 0 & & 0 & & 0 & &
\end{array}$$

prove that the top row is an exact sequence.

10.30 Prove that homology commutes with direct sums: For all n, there are natural isomorphisms

$$H_n\Big(\sum_\alpha \mathbf{C}^\alpha_\bullet\Big) \cong \sum_\alpha H_n(\mathbf{C}^\alpha_\bullet).$$

10.31 (i) Define a direct system of complexes $\{\mathbf{C}_\bullet^i, \varphi_j^i\}$, and prove that $\varinjlim \mathbf{C}_\bullet^i$ exists.

(ii) If $\{\mathbf{C}_\bullet^i, \varphi_j^i\}$ is a direct system of complexes over a directed index set, prove, for all $n \geq 0$, that

$$H_n(\varinjlim \mathbf{C}_\bullet^i) \cong \varinjlim H_n(\mathbf{C}_\bullet^i).$$

10.32 Suppose that a complex $(\mathbf{C}_\bullet, d_\bullet)$ of R-modules has a contracting homotopy in which the maps $s_n : C_n \to C_{n+1}$ satisfying

$$1_{C_n} = d_{n+1}s_n + s_{n-1}d_n$$

are only $\mathbb{Z}$-maps. Prove that $(\mathbf{C}_\bullet, d_\bullet)$ is an exact sequence.

10.33 (i) Let $0 \to F_n \to F_{n-1} \to \cdots \to F_0 \to 0$ be an exact sequence of finitely generated free k-modules, where k is a commutative ring. Prove that

$$\sum_{i=0}^{n} (-1)^i \operatorname{rank}(F_i) = 0.$$

(ii) Let

$$0 \to F_n \to F_{n-1} \to \cdots \to F_0 \to M \to 0$$

and

$$0 \to F'_m \to F'_{m-1} \to \cdots \to F'_0 \to M \to 0$$

be free resolutions of a k-module M in which all F_i and F'_j are finitely generated free k-modules. Prove that

$$\sum_{i=0}^{n} (-1)^i \operatorname{rank}(F_i) = \sum_{j=0}^{m} (-1)^j \operatorname{rank}(F'_j).$$

The common value is denoted by $\chi(M)$, and it is called the ***Euler–Poincaré characteristic*** of M.

Hint. Use Schanuel's lemma.

10.34 (i) Let $\mathbf{C}_\bullet : 0 \to C_n \to C_{n-1} \to \cdots \to C_0 \to 0$ be a complex of finitely generated free k-modules over a commutative ring k. Prove that

$$\sum_{i=0}^{n} (-1)^i \operatorname{rank}(C_i) = \sum_{i=0}^{n} (-1)^i \operatorname{rank}(H_i(\mathbf{C}_\bullet)).$$

(ii) Let $0 \to M' \to M \to M'' \to 0$ be an exact sequence of k-modules. If two of the modules have an Euler-Poincaré characteristic, prove that the third module does, too, and that

$$\chi(M) = \chi(M') + \chi(M'').$$

10.35 (i) (***Barratt–Whitehead***). Consider the commutative diagram with exact rows:

$$\begin{array}{ccccccccccc}
A_n & \xrightarrow{i_n} & B_n & \xrightarrow{p_n} & C_n & \xrightarrow{\partial_n} & A_{n-1} & \longrightarrow & B_{n-1} & \longrightarrow & C_{n-1} \\
\downarrow{\scriptstyle f_n} & & \downarrow{\scriptstyle g_n} & & \downarrow{\scriptstyle h_n} & & \downarrow{\scriptstyle f_{n-1}} & & \downarrow{\scriptstyle g_{n-1}} & & \downarrow{\scriptstyle h_{n-1}} \\
A'_n & \xrightarrow[j_n]{} & B'_n & \xrightarrow[q_n]{} & C'_n & \longrightarrow & A'_{n-1} & \longrightarrow & B'_{n-1} & \longrightarrow & C'_{n-1}
\end{array}$$

If each h_n is an isomorphism, prove that there is an exact sequence

$$A_n \xrightarrow{(f_n, i_n)} A'_n \oplus B_n \xrightarrow{j_n - g_n} B'_n \xrightarrow{\partial_n h_n^{-1} q_n} A_{n-1} \to A'_{n-1} \oplus B_{n-1} \to B'_{n-1},$$

where $(f_n, i_n)\colon a_n \mapsto (f_n a_n, i_n a_n)$ and $j_n - g_n\colon (a'_n, b_n) \mapsto j_n a'_n - g_n b_n$.

(ii) (***Mayer–Vietoris***). Assume, in the diagram of Theorem 10.44, that every third vertical map h_* is an isomorphism. Prove that there is an exact sequence

$$\cdots \to H_n(\mathbf{C}'_\bullet) \to H_n(\mathbf{A}'_\bullet) \oplus H_n(\mathbf{C}_\bullet) \to H_n(\mathbf{A}_\bullet) \to H_{n-1}(\mathbf{C}'_\bullet) \to \cdots.$$

Remark. The ***Eilenberg–Steenrod axioms*** characterize homology functors on the category **Top** of all topological spaces and continuous maps. If $h_n\colon \mathbf{Top} \to \mathbf{Ab}$ is a sequence of functors, for all $n \geq 0$, satisfying the long exact sequence, naturality of connecting homomorphisms, $h_n(f) = h_n(g)$ whenever f and g are homotopic, $h_0(X) = \mathbb{Z}$ and $h_n(X) = \{0\}$ for all $n > 0$ when X is a 1-point space, and *excision*, then there are natural isomorphisms $h_n \to H_n$ for all n. Now excision involves an added construction, called *relative homology*, but in the presence of the other axioms, excision can be replaced by exactness of the Mayer–Vietoris sequence. ◄

10.5 DERIVED FUNCTORS

In order to apply the general results about homology, we need a source of short exact sequences of complexes, as well as commutative diagrams in which they sit. The idea is to replace a module by a (deleted) resolution of it. We then apply either Hom or $\otimes$, and the resulting homology modules are called Ext or Tor. Given a short exact sequence of modules, we shall see that we may replace each of its modules by a resolution and obtain a short exact sequence of complexes.

This section is fairly dry, but it is necessary to establish the existence of homology functors. The most useful theorems in this section are Theorem 10.46 (the comparison theorem), Proposition 10.50 (which shows that the basic construction is well-defined), Corollary 10.57 (the long exact sequence), and Proposition 10.58 (naturality of the connecting homomorphism).

For those readers who are interested in using Tor (the left derived functors of tensor) and Ext (the right derived functors of Hom) immediately, and who are willing to defer looking at mazes of arrows, the next theorem gives a set of axioms characterizing the functors Ext^n.

Theorem 10.45. *Let* $\mathrm{EXT}^n\colon {}_R\mathbf{Mod} \to \mathbf{Ab}$ *be a sequence of contravariant functors, for* $n \geq 0$, *such that*

(i) *for every short exact sequence* $0 \to A \to B \to C \to 0$, *there is a long exact sequence and natural connecting homomorphisms*

$$\cdots \to \mathrm{EXT}^n(C) \to \mathrm{EXT}^n(B) \to \mathrm{EXT}^n(A) \xrightarrow{\Delta_n} \mathrm{EXT}^{n+1}(C) \to \cdots;$$

(ii) *there is a left R-module M with* EXT^0 *and* $\mathrm{Hom}_R(\ , M)$ *naturally equivalent*;

(iii) $\text{EXT}^n(P) = \{0\}$ *for all projective modules P and all $n \geq 1$.*

If $\text{Ext}^n(\ , M)$ *is another sequence of contravariant functors satisfying these same axioms, then* EXT^n *is naturally equivalent to* $\text{Ext}^n(\ , M)$ *for all $n \geq 0$.*

Remark. There are axiomatic descriptions of the covariant Ext functors and of the Tor functors in Exercises 10.44 and 10.45 on page 869. ◂

Proof. We proceed by induction on $n \geq 0$. The base step is axiom (ii).

For the inductive step, given a module A, choose a short exact sequence

$$0 \to L \to P \to A \to 0,$$

where P is projective. By axiom (i), there is a diagram with exact rows:

$$\begin{array}{ccccccc}
\text{EXT}^0(P) & \longrightarrow & \text{EXT}^0(L) & \xrightarrow{\Delta_0} & \text{EXT}^1(A) & \longrightarrow & \text{EXT}^1(P) \\
\downarrow{\scriptstyle \tau_P} & & \downarrow{\scriptstyle \tau_L} & & \vdots & & \\
\text{Hom}(P, M) & \longrightarrow & \text{Hom}(L, M) & \xrightarrow{\partial_0} & \text{Ext}^1(A, M) & \longrightarrow & \text{Ext}^1(P, M),
\end{array}$$

where the maps τ_P and τ_L are the isomorphisms given by axiom (ii). This diagram commutes because of the naturality of the equivalence $\text{EXT}^0 \to \text{Hom}(\ , M)$. By axiom (iii), $\text{Ext}^1(P, M) = \{0\}$ and $\text{EXT}^1(P) = \{0\}$. It follows that the maps Δ_0 and ∂_0 are surjective. This is precisely the sort of diagram in Proposition 8.93, and so there exists an isomorphism $\text{EXT}^1(A) \to \text{Ext}^1(A, M)$ making the augmented diagram commute.

We may now assume that $n \geq 1$, and we look further out in the long exact sequence. By axiom (i), there is a diagram with exact rows

$$\begin{array}{ccccccc}
\text{EXT}^n(P) & \longrightarrow & \text{EXT}^n(L) & \xrightarrow{\Delta_n} & \text{EXT}^{n+1}(A) & \longrightarrow & \text{EXT}^{n+1}(P) \\
 & & \downarrow{\scriptstyle \sigma} & & \vdots & & \\
\text{Ext}^n(P, M) & \longrightarrow & \text{Ext}^n(L, M) & \xrightarrow{\partial_n} & \text{Ext}^{n+1}(A, M) & \longrightarrow & \text{Ext}^{n+1}(P, M),
\end{array}$$

where $\sigma \colon \text{EXT}^n(L) \to \text{Ext}^n(L, M)$ is an isomorphism given by the inductive hypothesis. Since $n \geq 1$, all four terms involving the projective P are $\{0\}$; it follows from exactness of the rows that both Δ_n and ∂_n are isomorphisms. Finally, the composite $\partial_n \sigma \Delta_n^{-1} \colon \text{EXT}^{n+1}(A) \to \text{Ext}^{n+1}(A, M)$ is an isomorphism.

It remains to prove that the isomorphisms $\text{EXT}^n(A) \to \text{Ext}^n(A, M)$ constitute a natural transformation. It is here the assumed naturality in axiom (i) of the connecting homomorphism is used, and this is left for the reader to do. •

Such slow starting induction proofs, proving results for $n = 0$ and $n = 1$ before proving the inductive step, arise frequently, and they are called ***dimension shifting***.

The rest of this section consists of constructing functors that satisfy axioms (i), (ii), and (iii). We prove existence of Ext and Tor using derived functors (there are other proofs as well). As these functors are characterized by a short list of properties, we can usually work with Ext and Tor without being constantly aware of the details of their construction.

We begin with a technical definition.

Definition. If $\cdots \to P_2 \to P_1 \xrightarrow{d_1} P_0 \to A \to 0$ is a projective resolution of a module A, then its ***deleted projective resolution*** is the complex

$$\mathbf{P}_A = \cdots \to P_2 \to P_1 \to P_0 \to 0.$$

Similarly, if $0 \to A \to E^0 \xrightarrow{d^0} E^1 \to E^2 \to \cdots$ is an injective resolution of a module A, then a ***deleted injective resolution*** is the complex

$$\mathbf{E}^A = 0 \to E^0 \to E^1 \to E^2 \to \cdots.$$

In either case, deleting A loses no information: $A \cong \operatorname{coker} d_1$ in the first case, and $A \cong \ker d^0$ in the second case. Of course, a deleted resolution is no longer exact:

$$H_0(\mathbf{P}_A) = \ker(P_0 \to \{0\})/\operatorname{im} d_1 = P_0/\operatorname{im} d_1 \cong A.$$

We know that a module has many presentations, and so the next result is fundamental.

Theorem 10.46 (Comparison Theorem). *Given a map $f\colon A \to A'$, consider the diagram*

$$\begin{array}{ccccccccc}
\cdots \longrightarrow & P_2 & \xrightarrow{d_2} & P_1 & \xrightarrow{d_1} & P_0 & \xrightarrow{\varepsilon} & A & \longrightarrow 0 \\
 & \Big\downarrow{\scriptstyle \check{f}_2} & & \Big\downarrow{\scriptstyle \check{f}_1} & & \Big\downarrow{\scriptstyle \check{f}_0} & & \Big\downarrow{\scriptstyle f} & \\
\cdots \longrightarrow & P'_2 & \xrightarrow[d'_2]{} & P'_1 & \xrightarrow[d'_1]{} & P'_0 & \xrightarrow[\varepsilon']{} & A' & \longrightarrow 0,
\end{array}$$

where the rows are complexes. If each P_n in the top row is projective, and if the bottom row is exact, then there exists a chain map $\check{f}\colon \mathbf{P}_A \to \mathbf{P}'_{A'}$ making the completed diagram commute. Moreover, any two such chain maps are homotopic.

Remark. The dual of the comparison theorem is also true. Now the complexes go off to the right, the top row is assumed exact, and every term in the bottom row other than A' is injective. ◂

Proof. (i) We prove the existence of $\check{f}_n$ by induction on $n \geq 0$. For the base step $n = 0$, consider the diagram

$$\begin{array}{ccccc}
 & \swarrow & P_0 & & \\
 & & \Big\downarrow{\scriptstyle f\varepsilon} & & \\
P'_0 & \xrightarrow{\varepsilon'} & A' & \longrightarrow & 0
\end{array}$$

Since ε' is surjective and P_0 is projective, there exists a map $\check{f}_0 \colon P_0 \to P_0'$ with $\varepsilon' \check{f}_0 = f\varepsilon$.

For the inductive step, consider the diagram

$$\begin{array}{ccccc} P_{n+1} & \xrightarrow{d_{n+1}} & P_n & \xrightarrow{d_n} & P_{n-1} \\ & & \downarrow{\scriptstyle \check{f}_n} & & \downarrow{\scriptstyle \check{f}_{n-1}} \\ P'_{n+1} & \xrightarrow[d'_{n+1}]{} & P'_n & \xrightarrow[d'_n]{} & P'_{n-1} \end{array}$$

If we can show that im $\check{f}_n d_{n+1} \subseteq \operatorname{im} d'_{n+1}$, then we will have the diagram

$$\begin{array}{ccccc} & & P_{n+1} & & \\ & \swarrow & \downarrow{\scriptstyle \check{f}_n d_{n+1}} & & \\ P'_{n+1} & \xrightarrow{d'_{n+1}} & \operatorname{im} d'_{n+1} & \longrightarrow & 0 \end{array}$$

and projectivity of P_{n+1} will provide a map $\check{f}_{n+1} \colon P_{n+1} \to P'_{n+1}$ with $d'_{n+1}\check{f}_{n+1} = \check{f}_n d_{n+1}$. To check that the inclusion does hold, note that exactness at P'_n of the bottom row of the original diagram gives $\operatorname{im} d'_{n+1} = \ker d'_n$, and so it suffices to prove that $d'_n \check{f}_n d_{n+1} = 0$. But $d'_n \check{f}_n d_{n+1} = \check{f}_{n-1} d_n d_{n+1} = 0$.

(ii) We now prove uniqueness of $\check{f}$ to homotopy. If $h \colon \mathbf{P}_\bullet \to \mathbf{P}'_\bullet$ is a chain map also satisfying $\varepsilon' h_0 = f\varepsilon$, then we construct the terms $s_n \colon P_n \to P'_{n+1}$ of a homotopy s by induction on $n \geq 0$; that is, we want

$$\check{f}_n - h_n = d'_{n+1}s_n + s_{n-1}d_n.$$

Let us now begin the induction. If we define $s_0 = 0 = s_{-1}$, then $d'_1 s_0 + s_{-1}d_0 = 0$. On the other hand,

$$(\check{f}_0 - h_0)\varepsilon' = \check{f}_0\varepsilon' - h_0\varepsilon' = \varepsilon f - \varepsilon f = 0.$$

Since ε' is a surjection, we have $\check{f}_0 - h_0 = 0$, as desired.

For the inductive step, it suffices to prove that

$$\operatorname{im}(h_{n+1} - \check{f}_{n+1} - s_n d_{n+1}) \subseteq \operatorname{im} d'_{n+2},$$

for we have a diagram with exact row

$$\begin{array}{ccccc} & & P_{n+1} & & \\ & \swarrow & \downarrow{\scriptstyle h_{n+1} - \check{f}_{n+1} - s_n d_{n+1}} & & \\ P'_{n+2} & \xrightarrow{d'_{n+2}} & \operatorname{im} d'_{n+2} & \longrightarrow & 0 \end{array}$$

and projectivity of P_{n+1} will give a map $s_{n+1}\colon P_{n+1} \to P'_{n+2}$ satisfying the desired equation. As in the proof of part (i), exactness of the bottom row of the original diagram gives $\operatorname{im} d'_{n+2} = \ker d'_{n+1}$, and so it suffices to prove

$$d'_{n+1}(h_{n+1} - \check{f}_{n+1} - s_n d_{n+1}) = 0.$$

But

$$\begin{aligned} d'_{n+1}(h_{n+1} - \check{f}_{n+1} - s_n d_{n+1}) &= d'_{n+1}(h_{n+1} - \check{f}_{n+1}) - d'_{n+1} s_n d_{n+1} \\ &= d'_{n+1}(h_{n+1} - \check{f}_{n+1}) - (h_n - \check{f}_n - s_{n-1} d_n) d_{n+1} \\ &= d'_{n+1}(h_{n+1} - \check{f}_{n+1}) - (h_n - \check{f}_n) d_{n+1}, \end{aligned}$$

and the last term is 0 because h and $\check{f}$ are chain maps. •

We introduce a term to describe the chain map $\check{f}$ just constructed.

Definition. If $f\colon A \to A'$ is a map of modules, and if $\mathbf{P}_A$ and $\mathbf{P}'_{A'}$ are deleted projective resolutions of A and A', respectively, then a chain map $\check{f}\colon \mathbf{P}_A \to \mathbf{P}'_{A'}$

$$\begin{array}{ccccccccccc} \cdots & \longrightarrow & P_2 & \xrightarrow{d_2} & P_1 & \xrightarrow{d_1} & P_0 & \xrightarrow{\varepsilon} & A & \longrightarrow & 0 \\ & & \downarrow{\scriptstyle \check{f}_2} & & \downarrow{\scriptstyle \check{f}_1} & & \downarrow{\scriptstyle \check{f}_0} & & \downarrow{\scriptstyle f} & & \\ \cdots & \longrightarrow & P'_2 & \xrightarrow[d'_2]{} & P'_1 & \xrightarrow[d'_1]{} & P'_0 & \xrightarrow[\varepsilon']{} & A' & \longrightarrow & 0 \end{array}$$

is said to be ***over*** f if

$$f\varepsilon = \varepsilon' \check{f}_0.$$

Thus, the comparison theorem implies, given a homomorphism $f\colon A \to A'$, that a chain map over f always exists between deleted projective resolutions of A and A'; moreover, such a chain map is unique to homotopy.

Given a pair of rings R and S and an additive covariant functor $T\colon {}_R\mathbf{Mod} \to {}_S\mathbf{Mod}$, we are now going to construct, for all $n \in \mathbb{Z}$, its ***left derived functors*** $L_nT\colon {}_R\mathbf{Mod} \to {}_S\mathbf{Mod}$. The definition will be in two parts: first on objects; then on morphisms.

Choose, once for all, a deleted projective resolution $\mathbf{P}_A$ of every module A. As in Example 10.34(ix), form the complex $T\mathbf{P}_A$, and take homology:

$$L_nT(A) = H_n(T\mathbf{P}_A).$$

This definition is suggested by two examples. First, in algebraic topology, we tensor the complex of a triangulated space X to get homology groups $H_n(X; G)$ of X with *coefficients* in an abelian group G; or, we apply $\operatorname{Hom}(\ , G)$ to get a complex whose homology groups are called *cohomology groups* of X with coefficients in G (of course, this last functor is contravariant). Second, when we considered group extensions, the formulas that

arose suggested constructing a free resolution of the trivial module $\mathbb{Z}$, and then applying $\mathrm{Hom}(\ ,K)$ or $\otimes K$ to this resolution.

We now define $L_nT(f)$, where $f\colon A \to A'$ is a homomorphism. By the comparison theorem, there is a chain map $\check{f}\colon \mathbf{P}_A \to \mathbf{P}'_{A'}$ over f. It follows that $T\check{f}\colon T\mathbf{P}_A \to T\mathbf{P}'_{A'}$ is also a chain map, and we define $L_nT(f)\colon L_nT(A) \to L_nT(A')$ by

$$L_nT(f) = H_n(T\check{f}) = (T\check{f})_*.$$

In more detail, if $z \in \ker Td_n$, then

$$(L_nT)f\colon z + \mathrm{im}\, Td_{n+1} \mapsto (T\check{f}_n)z + \mathrm{im}\, Td'_{n+1}.$$

In pictures, look at the chosen projective resolutions:

$$\begin{array}{ccccccccc}
\cdots & \longrightarrow & P_1 & \longrightarrow & P_0 & \longrightarrow & A & \longrightarrow & 0 \\
 & & & & & & \big\downarrow{\scriptstyle f} & & \\
\cdots & \longrightarrow & P'_1 & \longrightarrow & P'_0 & \longrightarrow & A' & \longrightarrow & 0
\end{array}$$

Fill in the a chain map $\check{f}$ over f, then apply T to this diagram, and then take the map induced by $T\check{f}$ in homology.

Example 10.47.
If $r \in Z(R)$ is a central element in a ring R, and if A is a left R-module, then $\mu_r\colon A \to A$, defined by $\mu_r\colon A \mapsto rA$, is an R-map. We call μ_r *multiplication by r*.

Definition. A functor $T\colon {}_R\mathbf{Mod} \to {}_R\mathbf{Mod}$, of either variance, ***preserves multiplications*** if $T(\mu_r)\colon TA \to TA$ is multiplication by r for all $r \in Z(R)$.

Tensor product and Hom preserve multiplications. We claim that if T preserves multiplications, then L_nT also preserves multiplications; that is,

$$L_nT(\mu_r) = \text{ multiplication by } r.$$

Given a projective resolution $\cdots \to P_1 \xrightarrow{d_1} P_0 \xrightarrow{\varepsilon} A \to 0$, it is easy to see that $\check{\mu}$ is a chain map over μ_r,

$$\begin{array}{ccccccccccc}
\cdots & \longrightarrow & P_2 & \xrightarrow{d_2} & P_1 & \xrightarrow{d_1} & P_0 & \xrightarrow{\varepsilon} & A & \longrightarrow & 0 \\
 & & \big\downarrow{\scriptstyle \check{\mu}_2} & & \big\downarrow{\scriptstyle \check{\mu}_1} & & \big\downarrow{\scriptstyle \check{\mu}_0} & & \big\downarrow{\scriptstyle f} & & \\
\cdots & \longrightarrow & P_2 & \xrightarrow[d_2]{} & P_1 & \xrightarrow[d_1]{} & P_0 & \xrightarrow[\varepsilon]{} & A & \longrightarrow & 0,
\end{array}$$

where $\check{\mu}_n\colon P_n \to P_n$ is multiplication by r for every $n \ge 0$.. Since T preserves multiplications, the terms of the chain map $T\check{\mu}$ are multiplication by r, and so the induced maps in homology are also multiplication by r:

$$(T\check{\mu})_*\colon z_n + \mathrm{im}\, Td_{n+1} \mapsto (T\check{\mu}_n)z_n + \mathrm{im}\, Td_{n+1} = rz_n + \mathrm{im}\, Td_{n+1},$$

where $z_n \in \ker Td_n$. ◀

Proposition 10.48. *Given a pair of rings R and S and an additive covariant functor $T \colon {}_R\mathbf{Mod} \to {}_S\mathbf{Mod}$, then*

$$L_n T \colon {}_R\mathbf{Mod} \to {}_S\mathbf{Mod}$$

is an additive covariant functor for every n.

Proof. We will prove that $L_n T$ is well-defined on morphisms; it is then routine to check that it is a covariant additive functor (remember that H_n is a covariant additive functor from complexes to modules).

If $h \colon \mathbf{P}_A \to \mathbf{P}'_{A'}$ is another chain map over f, then the comparison theorem says that $h \simeq \check{f}$; therefore, $Th \simeq T\check{f}$, by Exercise 10.26 on page 828, and so $H_n(Th) = H_n(T\check{f})$, by Proposition 10.39. •

Proposition 10.49. *If $T \colon {}_R\mathbf{Mod} \to {}_S\mathbf{Mod}$ is a covariant additive functor, then $L_n T A = \{0\}$ for all negative n and for all A.*

Proof. By Exercise 10.21 on page 827, we have $L_n T A = \{0\}$ because, when n is negative, the nth term of $\mathbf{P}_A$ is $\{0\}$. •

Definition. If B is a left R-module and $T = \;\otimes_R B$, define

$$\mathrm{Tor}_n^R(\;, B) = L_n T.$$

Thus, if

$$\mathbf{P}_A = \cdots \to P_2 \xrightarrow{d_2} P_1 \xrightarrow{d_1} P_0 \to 0$$

is the chosen deleted projective resolution of a module A, then

$$\mathrm{Tor}_n^R(A, B) = H_n(\mathbf{P}_A \otimes_R B) = \frac{\ker(d_n \otimes 1_B)}{\mathrm{im}(d_{n+1} \otimes 1_B)}.$$

The domain of $\mathrm{Tor}_n^R(\;, B)$ is $\mathbf{Mod}_R$, the category of right R-modules; its target is $\mathbf{Ab}$, the category of abelian groups. For example, if R is commutative, then $A \otimes_R B$ is an R-module, and so the values of $\mathrm{Tor}_R(\;, B)$ lie in ${}_R\mathbf{Mod}$.

Definition. If A is a right R-module and $T = A \otimes_R \;$, define $\mathrm{tor}_n^R(A, \;) = L_n T$. Thus, if

$$\mathbf{Q}_B = \cdots \to Q_2 \xrightarrow{d_2} Q_1 \xrightarrow{d_1} Q_0 \to 0$$

is the chosen deleted projective resolution of a module B, then

$$\mathrm{tor}_n^R(A, B) = H_n(A \otimes_R \mathbf{Q}_B) = \frac{\ker(1_A \otimes d_n)}{\mathrm{im}(1_A \otimes d_{n+1})}.$$

The domain of $\mathrm{tor}_n^R(A, \;)$ is ${}_R\mathbf{Mod}$, the category of left R-modules; its target is $\mathbf{Ab}$, the category of abelian groups but, as before, its target may be smaller (if $R = \mathbb{Q}$, for example) or larger [if $R = \mathbb{Z}G$, for every $\mathbb{Z}$-module can be viewed as a (trivial) R-module].

One of the nice theorems of homological algebra is, for all A and B (and for all R and n), that

$$\mathrm{Tor}_n^R(A, B) \cong \mathrm{tor}_n^R(A, B).$$

There is a proof using spectral sequences, but there is also an elementary proof due to A. Zaks (see Rotman, *An Introduction to Homological Algebra*, p. 197).

There are now several points to discuss. The definition of $L_n T$ assumes that a choice of deleted projection resolution of each module has been made. Does $L_n T$ depend on this choice? And, once we dispose of this question (the answer is that $L_n T$ does not depend on the choice), how can we use these functors?

Assume that new choices $\widetilde{\mathbf{P}}_A$ of deleted projective resolutions have been made, and let us denote the left derived functors arising from these new choices by $\widetilde{L}_n T$.

Proposition 10.50. *Given a pair of rings R and S, and an additive covariant functor $T\colon {}_R\mathbf{Mod} \to {}_S\mathbf{Mod}$, then, for each n, the functors $L_n T$ and $\widetilde{L}_n T$ are naturally equivalent. In particular, for all A,*

$$(L_n T)A \cong (\widetilde{L}_n T)A,$$

and so these modules are independent of the choice of (*deleted*) *projective resolution of A.*

Proof. Consider the diagram

$$\begin{array}{ccccccccccc}
\cdots & \longrightarrow & P_2 & \longrightarrow & P_1 & \longrightarrow & P_0 & \longrightarrow & A & \longrightarrow & 0 \\
 & & & & & & & & \downarrow{\scriptstyle 1_A} & & \\
\cdots & \longrightarrow & \widetilde{P}_2 & \longrightarrow & \widetilde{P}_1 & \longrightarrow & \widetilde{P}_0 & \longrightarrow & A & \longrightarrow & 0,
\end{array}$$

where the top row is the chosen projective resolution of A used to define $L_n T$ and the bottom is that used to define $\widetilde{L}_n T$. By the comparison theorem, there is a chain map $\iota\colon \mathbf{P}_A \to \widetilde{\mathbf{P}}_A$ over 1_A. Applying T gives a chain map $T\iota\colon T\mathbf{P}_A \to T\widetilde{\mathbf{P}}_A$ over $T1_A = 1_{TA}$. This last chain map induces homomorphisms, one for each n,

$$\tau_A = (T\iota)_*\colon (L_n T)A \to (\widetilde{L}_n T)A.$$

We now prove that each τ_A is an isomorphism (thereby proving the last statement in the theorem) by constructing its inverse. Turn the preceding diagram upside down, so that the chosen projective resolution $\mathbf{P}_A \to A \to 0$ is now the bottom row. Again, the comparison theorem gives a chain map, say, $\kappa\colon \widetilde{\mathbf{P}}_A \to \mathbf{P}_A$. Now the composite $\kappa\iota$ is a chain map from $\mathbf{P}_A$ to itself over $1_{\mathbf{P_A}}$. By the uniqueness statement in the comparison theorem, $\kappa\iota \simeq 1_{\mathbf{P_A}}$; similarly, $\iota\kappa \simeq 1_{\widetilde{\mathbf{P}}_A}$. It follows that $T(\iota\kappa) \simeq 1_{T\widetilde{\mathbf{P}}_A}$ and $T(\kappa\iota) \simeq 1_{T\widetilde{\mathbf{P}}_A}$. Hence, $1 = (T\iota\kappa)_* = (T\iota)_*(T\kappa)_*$ and $1 = (T\kappa\iota)_* = (T\kappa)_*(T\iota)_*$. Therefore, $\tau_A = (T\iota)_*$ is an isomorphism.

We now prove that the isomorphisms τ_A constitute a natural equivalence: that is, if $f: A \to B$ is a homomorphism, then the following diagram commutes.

$$\begin{array}{ccc} (L_nT)A & \xrightarrow{\tau_A} & (\widetilde{L}_nT)A \\ {\scriptstyle L_nT(f)}\big\downarrow & & \big\downarrow{\scriptstyle \widetilde{L}_nT(f)} \\ (L_nT)B & \xrightarrow{\tau_B} & (\widetilde{L}_nT)B \end{array}$$

To evaluate in the clockwise direction, consider

$$\begin{array}{ccccccccc} \cdots & \longrightarrow & P_1 & \longrightarrow & P_0 & \longrightarrow & A & \longrightarrow & 0 \\ & & & & & & \big\downarrow{\scriptstyle 1_A} & & \\ \cdots & \longrightarrow & \widetilde{P}_1 & \longrightarrow & \widetilde{P}_0 & \longrightarrow & A & \longrightarrow & 0 \\ & & & & & & \big\downarrow{\scriptstyle f} & & \\ \cdots & \longrightarrow & \widetilde{Q}_1 & \longrightarrow & \widetilde{Q}_0 & \longrightarrow & B & \longrightarrow & 0, \end{array}$$

where the bottom row is some projective resolution of B. The comparison theorem gives a chain map $\mathbf{P}_A \to \widetilde{\mathbf{P}}_A$ over $f1_A = f$. Going counterclockwise, the picture will now have the chosen projective resolution of B as its middle row, and we get a chain map $\mathbf{P}_A \to \widetilde{\mathbf{P}}_A$ over $1_B f = f$. The uniqueness statement in the comparison theorem tells us that these two chain maps are homotopic, so that they give the same homomorphism in homology. Thus, the appropriate diagram commutes, showing that $\tau: L_nT \to \widetilde{L}_nT$ is a natural equivalence. •

Corollary 10.51. *The module* $\mathrm{Tor}_n^R(A, B)$ *is independent of the choices of projective resolutions of A and of B.*

Proof. The proposition applies at once to the left derived functors of $\quad\otimes_R B$, namely, $\mathrm{Tor}_n^R(\quad, B)$, and to the left derived functors of $A \otimes_R \quad$, namely $\mathrm{tor}_n^R(A, \quad)$. But we have already cited the fact that $\mathrm{Tor}_n^R(A, B) \cong \mathrm{tor}_n^R(A, B)$. •

Corollary 10.52. *Let* $T: {}_R\mathbf{Mod} \to {}_S\mathbf{Mod}$ *be an additive covariant functor. If P is a projective module, then* $L_nT(P) = \{0\}$ *for all* $n \geq 1$.

In particular, if A and P are right R-modules, with P projective, and if B and Q are left R-modules, with Q projective, then

$$\mathrm{Tor}_n^R(P, B) = \{0\} \quad \textit{and} \quad \mathrm{Tor}_n^R(A, Q) = \{0\}$$

for all $n \geq 1$.

Proof. Since P is projective, a projective resolution of it is

$$\mathbf{C}_\bullet = \cdots \to 0 \to 0 \to P \xrightarrow{1_P} P \to 0,$$

and so the corresponding deleted projective resolution $\mathbf{C}_P$ has only one nonzero term, namely, $C_0 = P$. It follows that $T\mathbf{C}_P$ is a complex having nth term $\{0\}$ for all $n \geq 1$, and so $L_nTP = H_n(T\mathbf{C}_P) = \{0\}$ for all $n \geq 1$, by Exercise 10.21 on page 827. •

We are now going to show that there is a long exact sequence of left derived functors. We begin with a useful lemma; it says that if we are given an exact sequence of modules as well as a projective resolution of its first and third terms, then we can "fill in the horseshoe"; that is, there is a projective resolution of the middle term that fits in the middle.

Lemma 10.53 (Horseshoe Lemma). *Given a diagram*

$$\begin{array}{ccccccccc}
 & & \downarrow & & & & \downarrow & & \\
 & & P_1' & & & & P_1'' & & \\
 & & \downarrow & & & & \downarrow & & \\
 & & P_0' & & & & P_0'' & & \\
 & & \downarrow{\scriptstyle \varepsilon'} & & & & \downarrow{\scriptstyle \varepsilon''} & & \\
0 & \longrightarrow & A' & \xrightarrow{i} & A & \xrightarrow{p} & A'' & \longrightarrow & 0,
\end{array}$$

where the columns are projective resolutions and the row is exact, then there exists a projective resolution of A and chain maps so that the three columns form an exact sequence of complexes.

Remark. The dual theorem, in which projective resolutions are replaced by injective resolutions, is also true. ◂

Proof. We show first that there is a projective P_0 and a commutative 3×3 diagram with exact columns and rows:

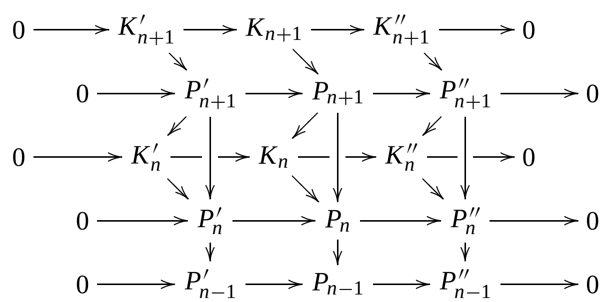
$$\begin{array}{ccccccccc}
 & & 0 & & 0 & & 0 & & \\
 & & \downarrow & & \downarrow & & \downarrow & & \\
0 & \longrightarrow & K_0' & \longrightarrow & K_0 & \longrightarrow & K_0'' & \longrightarrow & 0 \\
 & & \downarrow & & \downarrow & & \downarrow & & \\
0 & \longrightarrow & P_0' & \xrightarrow{i_0} & P_0 & \xrightarrow{p_0} & P_0'' & \longrightarrow & 0 \\
 & & \downarrow{\scriptstyle \varepsilon'} & & \downarrow{\scriptstyle \varepsilon} & & \downarrow{\scriptstyle \varepsilon''} & & \\
0 & \longrightarrow & A' & \xrightarrow{i} & A & \xrightarrow{p} & A'' & \longrightarrow & 0 \\
 & & \downarrow & & \downarrow & & \downarrow & & \\
 & & 0 & & 0 & & 0 & &
\end{array}$$

Define $P_0 = P_0' \oplus P_0''$; it is projective because both P_0' and P_0'' are projective. Define $i_0 \colon P_0' \to P_0' \oplus P_0''$ by $x' \mapsto (x', 0)$, and define $p_0 \colon P_0' \oplus P_0'' \to P_0''$ by $(x, x'') \mapsto x''$. It is clear that

$$0 \to P_0' \xrightarrow{i_0} P_0 \xrightarrow{p_0} P_0'' \to 0$$

is exact. Since P_0'' is projective, there exists a map $\sigma \colon P_0'' \to A$ with $p\sigma = \varepsilon''$. Now define $\varepsilon \colon P_0 \to A$ by $\varepsilon \colon (x', x'') \mapsto i\varepsilon' x' + \sigma x''$. It is left as a routine exercise that if $K_0 = \ker \varepsilon$, then there are maps $K_0' \to K_0$ and $K_0 \to K_0''$ (where $K_0' = \ker \varepsilon'$ and $K_0'' = \ker \varepsilon''$), so that the resulting 3×3 diagram commutes. Exactness of the top row is Exercise 10.29 on page 828.

We now prove, by induction on $n \geq 0$, that the bottom n rows of the desired diagram can be constructed. For the inductive step, assume that the first n steps have been filled in, and let $K_n = \ker(P_n \to P_{n-1})$, etc. Now construct the 3×3 diagram whose bottom row is $0 \to K_n' \to K_n \to K_n'' \to 0$, and splice it to the nth diagram, as illustrated next (note that the map $P_{n+1} \to P_n$ is defined as the composite $P_{n+1} \to K_n \to P_n$).



The columns of the new diagram are exact because, for example, $\operatorname{im}(P_{n+1} \to P_n) = K_n = \ker(P_n \to P_{n-1})$. •

Theorem 10.54. *If $0 \to A' \xrightarrow{i} A \xrightarrow{p} A'' \to 0$ is an exact sequence of modules and if $T \colon {}_R\mathbf{Mod} \to {}_S\mathbf{Mod}$ is a covariant additive functor, then there is a long exact sequence:*

$$\cdots \to L_n TA' \xrightarrow{L_n Ti} L_n TA \xrightarrow{L_n Tp} L_n TA'' \xrightarrow{\partial_n}$$
$$L_{n-1} TA' \xrightarrow{L_{n-1} Ti} L_{n-1} TA \xrightarrow{L_{n-1} Tp} L_{n-1} TA'' \xrightarrow{\partial_{n-1}} \cdots$$

that ends with

$$\cdots \to L_0 TA' \to L_0 TA \to L_0 TA'' \to 0.$$

Proof. Let $\mathbf{P'}_{A'}$ and $\mathbf{P}''_{A''}$ be the chosen deleted projective resolutions of A' and of A'', respectively. By Lemma 10.53, there is a deleted projective resolution $\widetilde{\mathbf{P}}_A$ of A with

$$\mathbf{0}_\bullet \to \mathbf{P'}_{A'} \xrightarrow{j} \widetilde{\mathbf{P}}_A \xrightarrow{q} \mathbf{P}''_{A''} \to \mathbf{0}_\bullet$$

(in the notation of the comparison theorem, $j = \check{i}$ is a chain map over i,and $q = \check{p}$ is a chain map over p). Applying T gives the sequence of complexes

$$\mathbf{0}_\bullet \to T\mathbf{P}'_{A'} \xrightarrow{Tj} T\widetilde{\mathbf{P}}_A \xrightarrow{Tq} T\mathbf{P}''_{A''} \to \mathbf{0}_\bullet.$$

To see that this sequence is exact,[12] note that each row $0 \to P'_n \xrightarrow{j_n} \widetilde{P}_n \xrightarrow{q_n} P''_n \to 0$ is a split exact sequence (because P''_n is projective), and additive functors preserve split short exact seqeunces. There is thus a long exact sequence

$$\cdots \to H_n(T\mathbf{P}'_{A'}) \xrightarrow{(Tj)_*} H_n(T\widetilde{\mathbf{P}}_A) \xrightarrow{(Tq)_*} H_n(T\mathbf{P}''_{A''}) \xrightarrow{\partial_n} H_{n-1}(T\mathbf{P}'_{A'}) \to \cdots;$$

that is, there is an exact sequence

$$\cdots \to L_nTA' \xrightarrow{(Ti)_*} \widetilde{L}_nTA \xrightarrow{(Tq)_*} L_nTA'' \xrightarrow{\partial_n} L_{n-1}TA' \to \cdots.$$

We do not know that the projective resolution of A given by the horseshoe lemma is the resolution originally chosen, and this is why we have $\widetilde{L}_nTA$ instead of L_nTA. But there is a natural equivalence $\tau\colon L_nT \to \widetilde{L}_nT$, and so there is an exact sequence

$$\cdots \to L_nTA' \xrightarrow{\tau_A^{-1}(Ti)_*} L_nTA \xrightarrow{\tau_A(Tq)_*} L_nTA'' \xrightarrow{\partial_n} L_{n-1}TA' \to \cdots.$$

The sequence does terminate with $\{0\}$, for $L_{-1}T = \{0\}$ for all negative n, by Proposition 10.49.

It remains to show that $\tau_A^{-1}(Ti)_* = L_nT(i)$ and $\tau_A^{-1}(Tq)_* = L_nT(p)$. Now $\tau_A^{-1} = (T\kappa)_*$, where $\kappa\colon \widetilde{\mathbf{P}}_A \to \mathbf{P}_A$ is a chain map over 1_A, and so

$$\tau_A^{-1}(Ti)_* = (T\kappa)_*(Ti)_* = \big(T(\kappa i)\big)_*.$$

Both κi and i are chain maps $\widetilde{\mathbf{P}}_A \to \mathbf{P}_A$ over 1_A, so they are homotopic, by the comparison theorem. Therefore, $T(\kappa i)$ and Ti are homotopic, and hence they induce the same map in homology: $(T(\kappa i))_* = (Ti)_* = L_nT(i)$, and so $\tau_A^{-1}(Ti)_* = L_nT(i)$. We prove $\tau_A^{-1}(Tq)_* = L_nT(p)$ in the same way. •

Corollary 10.55. *If $T\colon {}_R\mathbf{Mod} \to {}_S\mathbf{Mod}$ is a covariant additive functor, then the functor L_0T is right exact.*

Proof. This follows at once from the theorem. •

[12] The exact sequence of complexes is *not* split because the sequence of splitting maps need not constitute a chain map $\mathbf{P}''_{A''} \to \widetilde{\mathbf{P}}_A$.

Theorem 10.56.

(i) *If an additive covariant functor* $T\colon {}_R\mathbf{Mod} \to {}_S\mathbf{Mod}$ *is right exact, then* T *is naturally equivalent to* L_0T.

(ii) *The functor* $\;\otimes_R B$ *is naturally equivalent to* $\mathrm{Tor}_0^R(\;, B)$. *Hence, for all right* R*-modules* A*, there is an isomorphism*

$$A \otimes_R B \cong \mathrm{Tor}_0^R(A, B).$$

Proof. (i) Let $\mathbf{P}_A$ be the chosen deleted projective resolution of A, and let

$$\cdots \to P_1 \xrightarrow{d_1} P_0 \xrightarrow{\varepsilon} A \to 0$$

be the chosen projective resolution. By definition,

$$L_0TA = \frac{\ker(\varepsilon \otimes 1_B)}{\mathrm{im}(d_1 \otimes 1_B)} = \mathrm{coker}(d_1 \otimes 1_B).$$

But right exactness of T gives an exact sequence

$$TP_1 \xrightarrow{d_1 \otimes 1} TP_0 \xrightarrow{\varepsilon \otimes 1} TA \to 0.$$

Now $\varepsilon \otimes 1$ induces an isomorphism $\sigma_A\colon \mathrm{coker}(d_1 \otimes 1) \to TA$, by the first isomorphism theorem; that is, $\sigma_A\colon L_0TA \to TA$. It is left as a routine exercise that $\sigma\colon L_0T \to T$ is a natural equivalence.

(ii) Immediate from part (i), for $\;\otimes_R B$ is a covariant right exact additive functor. •

We have shown that Tor repairs the loss of exactness that may occur after tensoring a short exact sequence.

Corollary 10.57. *If* $0 \to A' \to A \to A'' \to 0$ *is a short exact sequence of modules, then there is a long exact sequence*

$$\begin{aligned}\cdots \to \mathrm{Tor}_2^R(A', B) \to \mathrm{Tor}_2^R(A, B) \to \mathrm{Tor}_2^R(A'', B)&\\ \to \mathrm{Tor}_1^R(A', B) \to \mathrm{Tor}_1^R(A, B) \to \mathrm{Tor}_1^R(A'', B)&\\ \to A' \otimes_R B \to A \otimes_R B \to A'' \otimes_R B \to 0.&\end{aligned}$$

The next proposition shows that the functors $\mathrm{Tor}_n(\;, B)$ satisfy the covariant version of Theorem 10.45.

Proposition 10.58. *Given a commutative diagram of modules having exact rows,*

$$\begin{array}{ccccccccc} 0 & \longrightarrow & A' & \xrightarrow{i} & A & \xrightarrow{p} & A'' & \longrightarrow & 0 \\ & & \downarrow{\scriptstyle f} & & \downarrow{\scriptstyle g} & & \downarrow{\scriptstyle h} & & \\ 0 & \longrightarrow & C' & \xrightarrow{j} & C & \xrightarrow{q} & C'' & \longrightarrow & 0 \end{array}$$

there is, for all n, a commutative diagram with exact rows

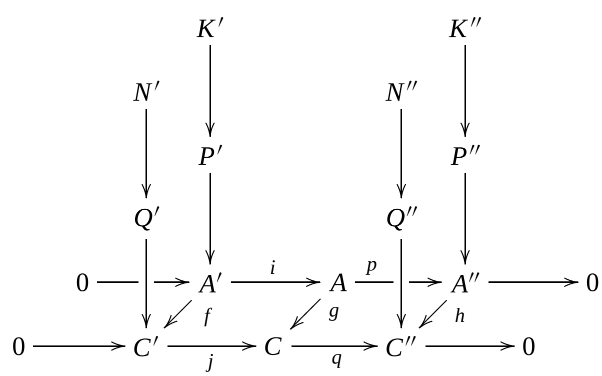
$$\begin{array}{ccccccc}
\mathrm{Tor}_n^R(A', B) & \xrightarrow{i_*} & \mathrm{Tor}_n^R(A, B) & \xrightarrow{p_*} & \mathrm{Tor}_n^R(A'', B) & \xrightarrow{\partial_n} & \mathrm{Tor}_{n-1}^R(A', B) \\
\downarrow{\scriptstyle f_*} & & \downarrow{\scriptstyle g_*} & & \downarrow{\scriptstyle h_*} & & \downarrow{\scriptstyle f_*} \\
\mathrm{Tor}_n^R(C', B) & \xrightarrow{j_*} & \mathrm{Tor}_n^R(C, B) & \xrightarrow{q_*} & \mathrm{Tor}_n^R(C'', B) & \xrightarrow{\partial'_n} & \mathrm{Tor}_{n-1}^R(C', B)
\end{array}$$

There is a similar diagram if the first variable is fixed.

Proof. Given the diagram in the statement, erect the chosen deleted projective resolutions on the corners $\mathbf{P}'_{A'}$, $\mathbf{P}''_{A''}$, $\mathbf{Q}'_{C'}$, and $\mathbf{Q}''_{C''}$. We claim that there are deleted projective resolutions $\widetilde{\mathbf{P}}_A$ and $\widetilde{\mathbf{Q}}_C$, together with chain maps, giving a commutative diagram of complexes having exact rows:

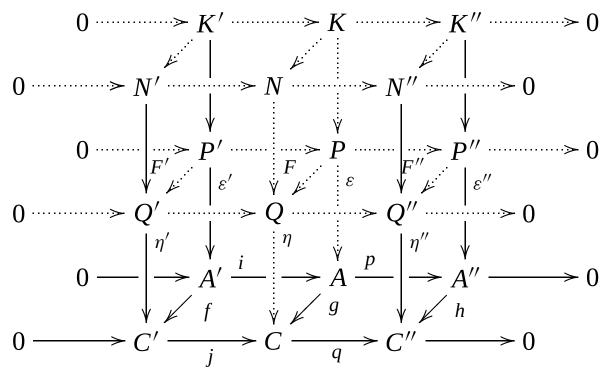
$$\begin{array}{ccccccccc}
0 & \longrightarrow & \mathbf{P}'_{A'} & \xrightarrow{\check{i}} & \widetilde{\mathbf{P}}_A & \xrightarrow{\check{p}} & \mathbf{P}''_{A''} & \longrightarrow & 0 \\
 & & \downarrow{\scriptstyle \check{f}} & & \downarrow{\scriptstyle \check{g}} & & \downarrow{\scriptstyle \check{h}} & & \\
0 & \longrightarrow & \mathbf{Q}'_{C'} & \xrightarrow{\check{j}} & \widetilde{\mathbf{Q}}_C & \xrightarrow{\check{q}} & \mathbf{Q}''_{C''} & \longrightarrow & 0
\end{array}$$

Once this is done, the result will follow from the naturality of the connecting homomorphism. As in the inductive proof of Theorem 10.44, it suffices to prove a three-dimensional version of the horseshoe lemma. We complete the following commutative diagram, whose columns are short exact sequences, and in which P', P'', Q', and Q'' are projectives and N', N'', K', and K'' are kernels,



to the following commutative diagram, whose rows and columns are short exact sequences,

and in which P and Q are projective:



Step 1. By the comparison theorem, there are chain maps $\check{f} : \mathbf{P}'_{A'} \to \mathbf{Q}'_{C'}$ over f and $\check{h} : \mathbf{P}''_{A''} \to \mathbf{Q}''_{C''}$ over h. To simplify notation, we will write $F' = \check{f}_0$ and $F'' = \check{h}_0$.

Step 2. Define $P = P' \oplus P''$, and insert the usual injection and projection maps $P' \to P$ and $P \to P''$, namely, $x' \mapsto (x', 0)$ and $(x', x'') \mapsto x''$. Similarly, define $Q = Q' \oplus Q''$, and insert the injection and projection maps $Q' \to Q$ and $Q \to Q''$. Of course, the sequences $0 \to P' \to P \to P'' \to 0$ and $0 \to Q' \to Q \to Q'' \to 0$ are exact.

Step 3. As in the proof of the horseshoe lemma, define $\varepsilon : P \to A$ by $\varepsilon : (x', x'') \mapsto i\varepsilon' x' + \sigma x''$, where $\sigma : P'' \to A$ satisfies $p\sigma = \varepsilon''$ (such a map σ was shown to exist in the proof of the horseshoe lemma); indeed, the horseshoe lemma shows that the rear face of the diagram commutes. Similarly, define $\eta : Q \to B$ by $\eta : (y', y'') \mapsto j\eta' y' + \tau y''$, where $\tau : Q'' \to B$ satisfies $q\tau = \eta''$; the front face commutes as well.

Step 4. Define $F : P \to Q$ by

$$F : (x', x'') \mapsto (F'x' + \gamma x'', F''x''),$$

where $\gamma : P'' \to Q'$ is to be constructed. It is easy to see that the plane containing the P's and Q's commutes, no matter how γ is defined.

Step 5. It remains to choose γ so that the square with vertices P, Q, C, and A commutes; that is, we want $f\varepsilon = \eta F$. Evaluating each side leads to the equation

$$fi\varepsilon' x' + f\sigma x'' = j\eta' F' x' + j\eta' \gamma x'' + \tau F'' x''.$$

Now $fi\varepsilon' = jf'\varepsilon' = j\eta' F'$ (because F' is the 0th term in the chain map $\check{f}$ over f), and so it suffices to find γ so that

$$j\varepsilon' \gamma = f\sigma - \tau F''.$$

Consider the diagram with exact row:

$$\begin{array}{ccccc} & & P'' & & \\ & & \big\downarrow{\scriptstyle f\sigma - \tau F''} & & \\ Q' & \xrightarrow{j\varepsilon'} & C & \xrightarrow{q} & C'' \end{array}$$

Now $\text{im}(f\sigma - \tau F'') \subseteq \text{im}\, j\varepsilon' = \ker q$, for

$$qf\sigma - q\tau F'' = f''p\sigma - \eta''F'' = f''\varepsilon'' - \eta''F'' = 0.$$

Since P'' is projective, there exists a map $\gamma : P'' \to Q'$ making the diagram commute.

Step 6. By the 3×3 Lemma (Exercise 10.29 on page 828), the rows $0 \to K' \to K \to K'' \to 0$ and $0 \to N' \to N \to N'' \to 0$ are exact, and we let the reader show that there are maps on the top face making every square commute. •

In the next section, we will show how Tor can be computed and used. But, before leaving this section, let us give the same treatment to Hom that we have just given to tensor product.

Left derived functors of a functor T are defined so that $T\mathbf{P}_A$ is a complex with all its nonzero terms on the *left* side; that is, all terms of negative degree are $\{0\}$. One consequence of this is Corollary 10.55: If T is right exact, then L_0T is naturally equivalent to T. As the Hom functors are left exact, we are now going to define right derived functors R^nT, in terms of deleted resolutions $\mathbf{C}_\bullet$ for which $T\mathbf{C}_\bullet$ is on the right. We shall see that R^0T is naturally equivalent to T when T is left exact.

Given an additive covariant functor $T : {}_R\mathbf{Mod} \to {}_S\mathbf{Mod}$, where R and S are rings, we are now going to construct, for all $n \in \mathbb{Z}$, its ***right derived functors*** $R^nT : {}_R\mathbf{Mod} \to {}_S\mathbf{Mod}$.

Choose, once for all, a deleted injective resolution $\mathbf{E}^A$ of every module A, form the complex $T\mathbf{E}^A$, and take homology:

$$R^nT(A) = H^n(T\mathbf{E}^A) = \frac{\ker Td^n}{\text{im}\, Td^{n-1}}.$$

The reader should reread Example 10.34(x) to recall the index raising convention; if the indices are lowered, then the definition would be

$$R^nT(A) = H_{-n}(T\mathbf{E}^A) = \frac{\ker Td_{-n}}{\text{im}\, Td_{-n+1}}.$$

Notice that we have raised the index on homology modules as well; we write H^n instead of H_{-n}.

The definition of $R^nT(f)$, where $f : A \to A'$ is a homomorphism, is similar to that for left derived functors. By the dual of the comparison theorem, there is a chain map

$\check{f}\colon \mathbf{E}^A \to \mathbf{E}'^{A'}$ over f, unique to homotopy, and so a unique map $R^nT(f)\colon H^n(T\mathbf{E}^A) \to H^n(T\mathbf{E}^{A'})$, namely, $(T\check{f}_n)_*$, is induced in homology.

In pictures, look at the chosen injective resolutions:

$$\begin{array}{ccccccccc} 0 & \longrightarrow & A' & \longrightarrow & E'^0 & \longrightarrow & E'^1 & \longrightarrow & \cdots \\ & & {\scriptstyle f}\big\uparrow & & & & & & \\ 0 & \longrightarrow & A & \longrightarrow & E^0 & \longrightarrow & E^1 & \longrightarrow & \cdots \end{array}$$

Fill in the a chain map $\check{f}$ over f, then apply T to this diagram, and then take the map induced by $T\check{f}$ in homology.

Proposition 10.59. *Given a pair of rings R and S and an additive covariant functor $T\colon {}_R\mathbf{Mod} \to {}_S\mathbf{Mod}$, then*

$$R^nT\colon {}_R\mathbf{Mod} \to {}_S\mathbf{Mod}$$

is an additive covariant functor for every n.

The proof of this proposition, as well as the proofs of other propositions about right derived functors soon to be stated, are essentially duals of the proofs we have already given, and so they will be omitted.

Example 10.60.
If T is a covariant additive functor that preserves multiplications, and if $\mu_r\colon A \to A$ is multiplication by r, where $r \in Z(R)$ is a central element, then R^nT also preserves multiplications (see Example 10.47). ◀

Proposition 10.61. *If $T\colon {}_R\mathbf{Mod} \to {}_S\mathbf{Mod}$ is a covariant additive functor, then $R^nTA = \{0\}$ for all negative n and for all A.*

Definition. If $T = \mathrm{Hom}_R(B, \ )$, define $\mathrm{Ext}^n_R(B, \ ) = R^nT$. Thus, if

$$\mathbf{E}^A = 0 \to E^0 \xrightarrow{d^0} E^1 \xrightarrow{d^1} E^2 \to \cdots,$$

is the chosen deleted injective resolution of a module A, then

$$\mathrm{Ext}^n_R(B, A) = H^n(\mathrm{Hom}_R(B, \mathbf{E}^A)) = \frac{\ker(d^n)_*}{\mathrm{im}(d^{n-1})_*},$$

where $(d^n)_*\colon \mathrm{Hom}_R(B, E^n) \to \mathrm{Hom}_R(B, E'^n)$ is defined, as usual, by

$$(d^n)_*\colon f \mapsto d^nf.$$

The domain of R^nT, in particular, the domain of $\mathrm{Ext}^n_R(B, \ )$, is ${}_R\mathbf{Mod}$, the category of all left R-modules, and its target is $\mathbf{Ab}$, the category of abelian groups. The target may be larger; for example, it is ${}_R\mathbf{Mod}$ if R is commutative.

Assume that new choices $\widetilde{\mathbf{E}}^A$ of deleted injective resolutions have been made, and let us denote the right derived functors arising from these new choices by $\widetilde{R}^nT$.

Proposition 10.62. *Given a pair of rings R and S, and an additive covariant functor $T\colon {}_R\mathbf{Mod} \to {}_S\mathbf{Mod}$, then, for each n, the functors R^nT and $\widetilde{R}^nT$ are naturally equivalent. In particular, for all A,*

$$(R^nT)A \cong (\widetilde{R}^nT)A,$$

and so these modules are independent of the choice of (deleted) injective resolution of A.

Corollary 10.63. *The module $\mathrm{Ext}^n_R(B, A)$ is independent of the choice of injective resolution of A.*

Corollary 10.64. *Let $T\colon {}_R\mathbf{Mod} \to {}_S\mathbf{Mod}$ be an additive covariant functor. If E is an injective module, then $R^nT(E) = \{0\}$ for all $n \geq 1$.*

In particular, if E is an injective R-module, then $\mathrm{Ext}^n_R(B, E) = \{0\}$ for all $n \geq 1$ and all modules B.

Theorem 10.65. *If $0 \to A' \xrightarrow{i} A \xrightarrow{p} A'' \to 0$ is an exact sequence of modules and if $T\colon {}_R\mathbf{Mod} \to {}_S\mathbf{Mod}$ is a covariant additive functor, then there is a long exact sequence:*

$$\cdots \to R^nTA' \xrightarrow{R^nTi} R^nTA \xrightarrow{R^nTp} R^nTA'' \xrightarrow{\partial^n}$$
$$R^{n+1}TA' \xrightarrow{R^{n+1}Ti} R^{n+1}TA \xrightarrow{R^{n+1}Tp} R^{n+1}TA'' \xrightarrow{\partial^{n+1}} \cdots$$

that begins with

$$0 \to R^0TA' \to R^0TA \to R^0TA'' \to \cdots.$$

Corollary 10.66. *If $T\colon {}_R\mathbf{Mod} \to {}_S\mathbf{Mod}$ is a covariant additive functor, then the functor R^0T is left exact.*

Theorem 10.67.

(i) *If an additive covariant functor $T\colon {}_R\mathbf{Mod} \to {}_S\mathbf{Mod}$ is left exact, then T is naturally equivalent to R^0T.*

(ii) *If B is a left R-module, the functor $\mathrm{Hom}_R(B, \ )$ is naturally equivalent to $\mathrm{Ext}^0_R(B, \ )$. Hence, for all left R-modules A, there is an isomorphism*

$$\mathrm{Hom}_R(B, A) \cong \mathrm{Ext}^0_R(B, A).$$

We have shown that Ext repairs the loss of exactness that may occur after applying Hom to a short exact sequence.

Corollary 10.68. *If $0 \to A' \to A \to A'' \to 0$ is a short exact sequence of modules, then there is a long exact sequence*

$$\begin{aligned} 0 \to \mathrm{Hom}_R(B, A') \to \mathrm{Hom}_R(B, A) \to \mathrm{Hom}_R(B, A'') & \\ \to \mathrm{Ext}^1_R(B, A') \to \mathrm{Ext}^1_R(B, A) \to \mathrm{Ext}^1_R(B, A'') & \\ \to \mathrm{Ext}^2_R(B, A') \to \mathrm{Ext}^2_R(B, A) \to \mathrm{Ext}^2_R(B, A'') \to \cdots. & \end{aligned}$$

Proposition 10.69. *Given a commutative diagram of modules having exact rows,*

$$\begin{array}{ccccccccc} 0 & \longrightarrow & A' & \xrightarrow{i} & A & \xrightarrow{p} & A'' & \longrightarrow & 0 \\ & & \downarrow{\scriptstyle f} & & \downarrow{\scriptstyle g} & & \downarrow{\scriptstyle h} & & \\ 0 & \longrightarrow & C' & \xrightarrow{j} & C & \xrightarrow{q} & C'' & \longrightarrow & 0 \end{array}$$

there is, for all n, a commutative diagram with exact rows

$$\begin{array}{ccccccc} \mathrm{Ext}^n_R(B, A') & \xrightarrow{i_*} & \mathrm{Ext}^n_R(B, A) & \xrightarrow{p_*} & \mathrm{Ext}^n_R(B, A'') & \xrightarrow{\partial^n} & \mathrm{Ext}^{n+1}_R(B, A') \\ \downarrow{\scriptstyle f_*} & & \downarrow{\scriptstyle g_*} & & \downarrow{\scriptstyle h_*} & & \downarrow{\scriptstyle f_*} \\ \mathrm{Ext}^n_R(B, C') & \xrightarrow{j_*} & \mathrm{Ext}^n_R(B, C) & \xrightarrow{q_*} & \mathrm{Ext}^n_R(B, C'') & \xrightarrow{\partial'^n} & \mathrm{Ext}^{n+1}_R(B, C') \end{array}$$

Finally, we discuss derived functors of contravariant functors T. If we define right derived functors $R^n T$, in terms of deleted resolutions $\mathbf{C}_\bullet$ for which $T\mathbf{C}_\bullet$ is on the right, then we start with a deleted projective resolution $\mathbf{P}_A$, for then the contravariance of T puts $T\mathbf{P}_A$ on the right.[13]

Given an additive contravariant functor $T\colon {}_R\mathbf{Mod} \to {}_S\mathbf{Mod}$, where R and S are rings, we are now going to construct, for all $n \in \mathbb{Z}$, its ***right derived functors*** $R^n T\colon {}_R\mathbf{Mod} \to {}_S\mathbf{Mod}$.

Choose, once for all, a deleted projective resolution $\mathbf{P}_A$ of every module A, form the complex $T\mathbf{P}_A$, and take homology:

$$R^n T(A) = H^n(T\mathbf{P}_A) = \frac{\ker T d_{n+1}}{\mathrm{im}\, T d_n}.$$

If $f\colon A \to A'$, define $R^n T(f)\colon R^n T(A') \to R^n T(A)$ as we did for left derived functors. By the comparison theorem, there is a chain map $\check{f}\colon \mathbf{P}_A \to \mathbf{P}'_{A'}$ over f, unique to homotopy, which induces a map $R^n T(f)\colon H^n(T\mathbf{P}'_{A'}) \to H^n(T\mathbf{P}_A)$, namely, $(T\check{f}_n)_*$, in homology.

[13] If we were interested in left derived functors of a contravariant T, but we are not, then we would use injective resolutions.

Example 10.70.
If T is an additive contravariant functor that preserves multiplications, and if $\mu_r : A \to A$ is multiplication by r, where $r \in Z(R)$ is a central element, then R^nT also preserves multiplications (see Example 10.47). ◂

Proposition 10.71. *Given a pair of rings R and S and an additive contravariant functor $T : {}_R\mathbf{Mod} \to {}_S\mathbf{Mod}$, then*

$$R^nT : {}_R\mathbf{Mod} \to {}_S\mathbf{Mod}$$

is an additive contravariant functor for every n.

Proposition 10.72. *If $T : {}_R\mathbf{Mod} \to {}_S\mathbf{Mod}$ is a contravariant additive functor, then $R^nTA = \{0\}$ for all negative n and for all A.*

Definition. If $T = \mathrm{Hom}_R(\ , C)$, define $\mathrm{ext}^n_R(\ , C) = R^nT$. Thus, if

$$\cdots \to P_2 \xrightarrow{d_2} P_1 \xrightarrow{d_1} P_0 \to 0$$

is the chosen deleted projective resolution of a module A, then

$$\mathrm{ext}^n_R(A, C) = H^n(\mathrm{Hom}_R(\mathbf{P}_A, C) = \frac{\ker(d_{n+1})^*}{\mathrm{im}(d_n)^*},$$

where $(d^n)^* : \mathrm{Hom}_R(P'_n, C) \to \mathrm{Hom}_R(P_n, C)$ is defined, as usual, by

$$(d_n)^* : f \mapsto fd_n.$$

The same phenomenon that holds for Tor holds for Ext: for all A and C (and for all R and n),

$$\mathrm{Ext}^n_R(A, C) \cong \mathrm{ext}^n_R(A, C).$$

The same proof that shows that Tor is independent of the variable resolved also works for Ext (see Rotman, *An Introduction to Homological Algebra*, p. 197). In light of this theorem, we will dispense with the two notations for Ext.

Assume that new choices $\widetilde{\mathbf{P}}_A$ of deleted projective resolutions have been made, and let us denote the right derived functors arising from these new choices by $\widetilde{R}^nT$.

Proposition 10.73. *Given a pair of rings R and S, and an additive contravariant functor $T : {}_R\mathbf{Mod} \to {}_S\mathbf{Mod}$, then, for each n, the functors R^nT and $\widetilde{R}^nT$ are naturally equivalent. In particular, for all A,*

$$(R^nT)A \cong (\widetilde{R}^nT)A,$$

and so these modules are independent of the choice of (deleted) projective resolution of A.

Corollary 10.74. *The module* $\operatorname{Ext}_R^n(A, C)$ *is independent of the choice of projective resolution of* A.

Corollary 10.75. *Let* $T\colon {}_R\mathbf{Mod} \to {}_S\mathbf{Mod}$ *be an additive contravariant functor. If* P *is a projective module, then* $R^nT(P) = \{0\}$ *for all* $n \geq 1$.

In particular, if P *is a projective* R*-module, then* $\operatorname{Ext}_R^n(P, B) = \{0\}$ *for all* $n \geq 1$ *and all modules* B.

Theorem 10.76. *If* $0 \to A' \xrightarrow{i} A \xrightarrow{p} A'' \to 0$ *is an exact sequence of modules and if* $T\colon {}_R\mathbf{Mod} \to {}_S\mathbf{Mod}$ *is a contravariant additive functor, then there is a long exact sequence*

$$\cdots \to R^nTA'' \xrightarrow{R^nTp} R^nTA \xrightarrow{R^nTi} R^nTA' \xrightarrow{\partial^n} R^{n+1}TA'' \xrightarrow{R^{n+1}Tp} R^{n+1}TA \xrightarrow{R^{n+1}Ti} R^{n+1}TA' \xrightarrow{\partial^{n+1}} \cdots$$

that begins with

$$0 \to R^0TA'' \to R^0TA \to R^0TA' \to \cdots.$$

Corollary 10.77. *If* $T\colon {}_R\mathbf{Mod} \to {}_S\mathbf{Mod}$ *is a contravariant additive functor, then the functor* R^0T *is left exact.*

Theorem 10.78.

(i) *If an additive contravariant functor* $T\colon {}_R\mathbf{Mod} \to {}_S\mathbf{Mod}$ *is left exact, then* T *is naturally equivalent to* R^0T.

(ii) *If* C *is a left* R*-module, the functor* $\operatorname{Hom}_R(\ , C)$ *is naturally equivalent to* $\operatorname{Ext}_R^0(\ , C)$. *Hence, for all left* R*-modules* A*, there is an isomorphism*

$$\operatorname{Hom}_R(A, C) \cong \operatorname{Ext}_R^0(A, C).$$

We have shown that Ext repairs the loss of exactness that may occur after applying Hom to a short exact sequence.

Corollary 10.79. *If* $0 \to A' \to A \to A'' \to 0$ *is a short exact sequence of modules, then there is a long exact sequence*

$$\begin{aligned} 0 &\to \operatorname{Hom}_R(A'', C) \to \operatorname{Hom}_R(A, C) \to \operatorname{Hom}_R(A', C) \\ &\to \operatorname{Ext}_R^1(A'', C) \to \operatorname{Ext}_R^1(A, C) \to \operatorname{Ext}_R^1(A', C) \\ &\to \operatorname{Ext}_R^2(A'', C) \to \operatorname{Ext}_R^2(A, C) \to \operatorname{Ext}_R^2(A', C) \to \cdots . \end{aligned}$$

Proposition 10.80. *Given a commutative diagram of modules having exact rows,*

$$\begin{array}{ccccccccc} 0 & \longrightarrow & A' & \xrightarrow{i} & A & \xrightarrow{p} & A'' & \longrightarrow & 0 \\ & & \downarrow{\scriptstyle f} & & \downarrow{\scriptstyle g} & & \downarrow{\scriptstyle h} & & \\ 0 & \longrightarrow & C' & \xrightarrow{j} & C & \xrightarrow{q} & C'' & \longrightarrow & 0 \end{array}$$

there is, for all n, a commutative diagram with exact rows

$$\begin{array}{ccccccc} \mathrm{Ext}^n_R(A'', B) & \xrightarrow{p^*} & \mathrm{Ext}^n_R(A, B) & \xrightarrow{i^*} & \mathrm{Ext}^n_R(A', B) & \xrightarrow{\partial^n} & \mathrm{Ext}^{n+1}_R(A'', B) \\ {\scriptstyle h^*}\uparrow & & \uparrow{\scriptstyle g^*} & & \uparrow{\scriptstyle f^*} & & \uparrow{\scriptstyle h^*} \\ \mathrm{Ext}^n_R(C'', B) & \xrightarrow{q^*} & \mathrm{Ext}^n_R(C, B) & \xrightarrow{j^*} & \mathrm{Ext}^n_R(C', B) & \xrightarrow{\partial^{n\prime}} & \mathrm{Ext}^{n+1}_R(C'', B) \end{array}$$

Remark. When T is a covariant functor, then we call the ingredients of L_nT chains, cycles, boundaries, and homology. When T is contravariant, we often add the prefix "co," and the ingredients of R^nT are usually called *cochains*, *cocycles*, *coboundaries*, and *cohomology*. Unfortunately, this clear distinction is blurred because the Hom functor is contravariant in one variable but covariant in the other. In spite of this, we usually use the "co" prefix for the derived functors Ext^n of Hom. ◄

Derived functors are one way to construct functors like Ext and Tor. In the next section, along with more properties of Ext and Tor, we shall describe another construction of Ext, due to N. Yoneda, and another construction of Tor, due to S. Mac Lane. Indeed, derived functors will rarely be mentioned in the sequel.

EXERCISES

10.36 If $\tau\colon F \to G$ is a natural transformation between additive functors, prove that τ gives chain maps $\tau_{\mathbf{C}}\colon F\mathbf{C} \to G\mathbf{C}$ for every complex $\mathbf{C}$. If τ is a natural equivalence, prove that $F\mathbf{C} \cong G\mathbf{C}$.

10.37 (i) Let $T\colon {}_R\mathbf{Mod} \to {}_S\mathbf{Mod}$ be an exact additive functor, where R and S are rings, and suppose that P projective implies TP projective. If B is a left R-module and $\mathbf{P}_B$ is a deleted projective resolution of B, prove that $T\mathbf{P}_{TB}$ is a deleted projective resolution of TB.

(ii) Let A be an R-algebra, where R is a commutative ring, which is flat as an R-module. Prove that if B is an A-module (and hence an R-module), then

$$A \otimes_R \mathrm{Tor}_n^R(B, C) \cong \mathrm{Tor}_n^A(B, A \otimes_R C)$$

for all R-modules C and all $n \geq 0$.

10.38 Let R be a semisimple ring.

(i) Prove, for all $n \geq 1$, that $\mathrm{Tor}_n^R(A, B) = \{0\}$ for all right R-modules A and all left R-modules B.

Hint. If R is semisimple, then every (left or right) R-module is projective.

(ii) Prove, for all $n \geq 1$, that $\mathrm{Ext}_R^n(A, B) = \{0\}$ for all left R-modules A and B.

10.39 If R is a PID, prove, for all $n \geq 2$, that $\mathrm{Tor}_n^R(A, B) = \{0\} = \mathrm{Ext}_R^n(A, B)$ for all R-modules A and B.
Hint. Use Theorem 9.8.

10.40 Let R be a domain and let A be an R-module.

(i) Prove that if the multiplication $\mu_r \colon A \to A$ is an injection for all $r \neq 0$, then A is torsion-free.

(ii) Prove that if the multiplication $\mu_r \colon A \to A$ is a surjection for all $r \neq 0$, then A is divisible.

(iii) Prove that if the multiplication $\mu_r \colon A \to A$ is an isomorphism for all $r \neq 0$, then A is a vector space over Q, where $Q = \mathrm{Frac}(R)$.

Hint. A module A is a vector space over Q if and only if it is torsion-free and divisible.

(iv) If either C or A is a vector space over Q, prove that $\mathrm{Tor}_n^R(C, A)$ and $\mathrm{Ext}_R^n(C, A)$ are also vector spaces over Q.

10.41 Let R be a domain and let $Q = \mathrm{Frac}(R)$.

(i) If $r \in R$ is nonzero and A is an R-module for which $rA = \{0\}$; that is, $ra = 0$ for all $a \in A$, prove that $\mathrm{Ext}_R^n(Q, A) = \{0\} = \mathrm{Tor}_n^R(Q, A)$ for all $n \geq 0$.

Hint. If V is a vector space over Q for which $rV = \{0\}$, then $V = \{0\}$.

(ii) Prove that $\mathrm{Ext}_R^n(V, A) = \{0\} = \mathrm{Tor}_n^R(V, A)$ for all $n \geq 0$ whenever V is a vector space over Q and A is an R-module for which $rA = \{0\}$ for some nonzero $r \in R$.

10.42 Let A and B be R-modules. For $f \colon A' \to B$, where A' is a submodule of A, define its ***obstruction*** to be $\partial(f)$, where $\partial \colon \mathrm{Hom}_R(A', B) \to \mathrm{Ext}_R^1(A/A', B)$ is the connecting homomorphism. Prove that f can be extended to a homomorphism $\widetilde{f} \colon A \to B$ if and only if its obstruction is 0.

10.43 If $T \colon \mathbf{Ab} \to \mathbf{Ab}$ is a left exact functor, prove that $L_0 T$ is an exact functor. Conclude, for any abelian group B, that $L_0 \mathrm{Hom}(B, \ )$ is not naturally equivalent to $\mathrm{Hom}(B, \ )$.

10.6 EXT AND TOR

We now examine Ext and Tor more closely. As we said in the last section, all properties of these functors should follow from versions of Theorem 10.45, the axioms characterizing them (see Exercises 10.44 and 10.45 on page 869); in particular, their construction as derived functors need not be used.

We begin by showing that Ext behaves like Hom with respect to sums and products.

Proposition 10.81. *If $\{A_k : k \in K\}$ is a family of modules, then there are natural isomorphisms, for all n,*

$$\operatorname{Ext}_R^n\Big(\sum_{k\in K} A_k, B\Big) \cong \prod_{k\in K} \operatorname{Ext}_R^n(A_k, B).$$

Proof. The proof is by dimension shifting; that is, by induction on $n \geq 0$. The base step is Theorem 7.33, for $\operatorname{Ext}^0(\ , B)$ is naturally equivalent to the contravariant functor $\operatorname{Hom}(\ , B)$.

For the inductive step, choose, for each $k \in K$, a short exact sequence

$$0 \to L_k \to P_k \to A_k \to 0,$$

where P_k is projective. There is an exact sequence

$$0 \to \sum_k L_k \to \sum_k P_k \to \sum_k A_k \to 0,$$

and $\sum_k P_k$ is projective, for every sum of projectives is projective. There is a commutative diagram with exact rows:

$$\begin{array}{ccccccc}
\operatorname{Hom}(\sum P_k, B) & \to & \operatorname{Hom}(\sum L_k, B) & \xrightarrow{\partial} & \operatorname{Ext}^1(\sum A_k, B) & \longrightarrow & \operatorname{Ext}^1(\sum P_k, B) \\
\downarrow{\scriptstyle \tau} & & \downarrow{\scriptstyle \sigma} & & \vdots & & \\
\prod \operatorname{Hom}(P_k, B) & \to & \prod \operatorname{Hom}(L_k, B) & \xrightarrow{d} & \prod \operatorname{Ext}^1(A_k, B) & \to & \prod \operatorname{Ext}^1(\sum P_k, B),
\end{array}$$

where the maps in the bottom row are just the usual induced maps in each coordinate, and the maps τ and σ are the isomorphisms given by Theorem 7.33. Now $\operatorname{Ext}^1(\sum P_k, B) = \{0\} = \prod \operatorname{Ext}^1(P_k, B)$, because $\sum P_k$ and each P_k are projective, so that the maps ∂ and d are surjective. This is precisely the sort of diagram in Proposition 8.93, and so there exists an isomorphism $\operatorname{Ext}^1(\sum A_k, B) \to \prod \operatorname{Ext}^1(A_k, B)$ making the augmented diagram commute.

We may now assume that $n \geq 1$, and we look further out in the long exact sequence. There is a commutative diagram

$$\begin{array}{ccccccc}
\operatorname{Ext}^n\big(\sum P_k, B\big) & \to & \operatorname{Ext}^n\big(\sum L_k, B\big) & \xrightarrow{\partial} & \operatorname{Ext}^{n+1}\big(\sum A_k, B\big) & \to & \operatorname{Ext}^{n+1}\big(\sum P_k, B\big) \\
 & & \downarrow{\scriptstyle \sigma} & & \vdots & & \\
\prod \operatorname{Ext}^n(P_k, B) & \to & \prod \operatorname{Ext}^n(L_k, B) & \xrightarrow{d} & \prod \operatorname{Ext}^{n+1}(A_k, B) & \to & \prod \operatorname{Ext}^{n+1}(P_k, B),
\end{array}$$

where $\sigma\colon \operatorname{Ext}^n(\sum L_k, B) \to \prod \operatorname{Ext}^n(L_k, B)$ is an isomorphism that exists by the inductive hypothesis. Since $n \geq 1$, all four Ext's whose first variable is projective are $\{0\}$; it follows from exactness of the rows that both ∂ and d are isomorphisms. Finally, the composite $d\sigma\partial^{-1}\colon \operatorname{Ext}^{n+1}(\sum A_k, B) \to \prod \operatorname{Ext}^{n+1}(A_k, B)$ is an isomorphism, as desired. •

There is a dual result in the second variable.

Proposition 10.82. *If $\{B_k : k \in K\}$ is a family of modules, then there are natural isomorphisms, for all n,*

$$\mathrm{Ext}^n_R\big(A, \prod_{k \in K} B_k\big) \cong \prod_{k \in K} \mathrm{Ext}^n_R(A, B_k).$$

Proof. The proof is by dimension shifting. The base step is Theorem 7.32, for $\mathrm{Ext}^0(A,\ )$ is naturally equivalent to the covariant functor $\mathrm{Hom}(A,\ )$.

For the inductive step, choose, for each $k \in K$, a short exact sequence

$$0 \to B_k \to E_k \to N_k \to 0,$$

where E_k is injective. There is an exact sequence

$$0 \to \prod_k B_k \to \prod_k E_k \to \prod_k N_k \to 0,$$

and $\prod_k E_k$ is injective, for every product of injectives is injective, by Proposition 7.66. There is a commutative diagram with exact rows:

$$\begin{array}{ccccccc}
\mathrm{Hom}(A, \prod E_k) & \to & \mathrm{Hom}(A, \prod N_k) & \xrightarrow{\partial} & \mathrm{Ext}^1(A, \prod B_k) & \to & \mathrm{Ext}^1(A, \prod E_k) \\
\downarrow{\scriptstyle \tau} & & \downarrow{\scriptstyle \sigma} & & \vdots & & \\
\prod \mathrm{Hom}(A, E_k) & \to & \prod \mathrm{Hom}(A, N_k) & \xrightarrow{d} & \prod \mathrm{Ext}^1(A, B_k) & \to & \prod \mathrm{Ext}^1(A, E_k),
\end{array}$$

where the maps in the bottom row are just the usual induced maps in each coordinate, and the maps τ and σ are the isomorphisms given by Theorem 7.32. The proof now finishes as that of Proposition 10.81. •

It follows that Ext^n commutes with finite direct sums in either variable.

Remark. These last two propositions cannot be generalized by replacing sums by direct limits or products by inverse limits; the reason is that direct limits of projectives need not be projective and inverse limits of injectives need not be injective. ◂

When the ring R is noncommutative, $\mathrm{Hom}_R(A, B)$ is an abelian group, but it need not be an R-module.

Proposition 10.83.

(i) *Let $r \in Z(R)$ be a central element, and let A and B be left R-modules. If $\mu_r : B \to B$ is multiplication by r, then the induced map*

$$\mu_r^* \colon \mathrm{Ext}^n_R(A, B) \to \mathrm{Ext}^n_R(A, B)$$

is also multiplication by r. A similar statement is true in the other variable.

(ii) *If R is a commutative ring, then $\mathrm{Ext}^n_R(A, B)$ is an R-module.*

Proof. (i) This follows at once from Example 10.47.

(ii) This follows from part (i) if we define scalar multiplication by r to be μ_r^*. •

Example 10.84.
(i) We show, for every abelian group B, that

$$\operatorname{Ext}^1_{\mathbb{Z}}(\mathbb{I}_n, B) \cong B/nB.$$

There is an exact sequence

$$0 \to \mathbb{Z} \xrightarrow{\mu_n} \mathbb{Z} \to \mathbb{I}_n \to 0,$$

where μ_n is multiplication by n. Applying $\operatorname{Hom}(\ , B)$ gives exactness of

$$\operatorname{Hom}(\mathbb{Z}, B) \xrightarrow{\mu_n^*} \operatorname{Hom}(\mathbb{Z}, B) \to \operatorname{Ext}^1(\mathbb{I}_n, B) \to \operatorname{Ext}^1(\mathbb{Z}, B).$$

Now $\operatorname{Ext}^1(\mathbb{Z}, B) = \{0\}$ because $\mathbb{Z}$ is projective. Moreover, μ_n^* is also multiplication by n, while $\operatorname{Hom}(\mathbb{Z}, B) = B$. More precisely, $\operatorname{Hom}(\mathbb{Z},\ )$ is naturally equivalent to the identity functor on **Ab**, and so there is a commutative diagram with exact rows

$$\begin{array}{ccccccc}
B & \xrightarrow{\mu_n} & B & \longrightarrow & B/nB & \longrightarrow & 0 \\
\downarrow{\scriptstyle \tau_B} & & \downarrow{\scriptstyle \tau_B} & & \vdots & & \\
\operatorname{Hom}(\mathbb{Z}, B) & \xrightarrow{\mu_n^*} & \operatorname{Hom}(\mathbb{Z}, B) & \longrightarrow & \operatorname{Ext}^1(\mathbb{I}_n, B) & \longrightarrow & 0
\end{array}$$

By Proposition 8.93, there is an isomorphism $B/nB \cong \operatorname{Ext}^1(\mathbb{I}_n, B)$.

(ii) We can now compute $\operatorname{Ext}^1_{\mathbb{Z}}(A, B)$ whenever A and B are finitely generated abelian groups. By the fundamental theorem, both A and B are direct sums of cyclic groups. Since Ext commutes with finite direct sums, $\operatorname{Ext}^1_{\mathbb{Z}}(A, B)$ is the direct sum of groups $\operatorname{Ext}^1_{\mathbb{Z}}(C, D)$, where C and D are cyclic. We may assume that C is finite, otherwise it is projective, and $\operatorname{Ext}^1(C, D) = \{0\}$. This calculation can be completed using part (i) and Exercise 5.5 on page 267, which says that if D is a cyclic group of finite order m, then D/nD is a cyclic group of order d, where $d = (m, n)$ is their gcd. ◄

We now give the obvious definition analogous to extensions of groups.

Definition. Given R-modules C and A, an ***extension*** of A by C is a short exact sequence

$$0 \to A \xrightarrow{i} B \xrightarrow{p} C \to 0.$$

An extension is ***split*** if there exists an R-map $s : C \to B$ with $ps = 1_C$.

Of course, if $0 \to A \to B \to C \to 0$ is a split extension, then $B \cong A \oplus C$.

Whenever meeting a homology group, we must ask what it means for it to be zero, for its elements can then be construed as being obstructions. For example, factor sets explain why a group extension may not be split. In this section, we will show that $\operatorname{Ext}^1_R(C, A) = \{0\}$ if and only if every extension of A by C splits. Thus, nonzero elements of any $\operatorname{Ext}^1_R(C', A')$ describe nonsplit extensions (indeed, this result is why Ext is so called) .

We begin with a definition motivated by Proposition 10.17.

Definition. Given modules C and A, two extensions $\xi : 0 \to A \to B \to C \to 0$ and $\xi' : 0 \to A \to B' \to C \to 0$ of A by C are ***equivalent*** if there exists a map $\varphi \colon B \to B'$ making the following diagram commute:

$$\begin{array}{ccccccccc}
\xi : 0 & \longrightarrow & A & \xrightarrow{i} & B & \xrightarrow{p} & C & \longrightarrow & 1 \\
 & & \downarrow{\scriptstyle 1_A} & & \downarrow{\scriptstyle \varphi} & & \downarrow{\scriptstyle 1_C} & & \\
\xi' : 0 & \longrightarrow & A & \xrightarrow{i'} & B' & \xrightarrow{p'} & C & \longrightarrow & 1
\end{array}$$

We denote the equivalence class of an extension ξ by $[\xi]$, and we define

$$e(C, A) = \big\{[\xi] : \xi \text{ is an extension of } A \text{ by } C\big\}.$$

If two extensions are equivalent, then the five lemma (Exercise 8.52 on page 604) shows that the map φ must be an isomorphism; it follows that equivalence is, indeed, an equivalence relation (for we can now prove symmetry). However, the converse is false: There can be inequivalent extensions having isomorphic middle terms, as we saw in Example 10.18 (all groups in this example are abelian, and so we may view it as an example of $\mathbb{Z}$-modules).

Proposition 10.85. *If* $\operatorname{Ext}^1_R(C, A) = \{0\}$, *then every extension*

$$0 \to A \xrightarrow{i} B \xrightarrow{p} C \to 0$$

is split.

Proof. Apply the functor $\operatorname{Hom}(C, \ )$ to the extension to obtain an exact sequence

$$\operatorname{Hom}(C, B) \xrightarrow{p_*} \operatorname{Hom}(C, C) \xrightarrow{\partial} \operatorname{Ext}^1(C, A).$$

By hypothesis, $\operatorname{Ext}^1(C, A) = \{0\}$, so that p_* is surjective. Hence, there exists $s \in \operatorname{Hom}(C, B)$ with $1_C = p_*(s)$; that is, $1_C = ps$, and this says that the extension splits. •

Corollary 10.86. *An R-module P is projective if and only if* $\operatorname{Ext}^1_R(P, B) = \{0\}$ *for every R-module B.*

Proof. If P is projective, then $\operatorname{Ext}^1_R(P, B) = \{0\}$ for all B, by Corollary 10.52. Conversely, if $\operatorname{Ext}^1_R(P, B) = \{0\}$ for all B, then every exact sequence $0 \to B \to X \to P \to 0$ splits, by Proposition 10.85, and so P is projective, by Proposition 7.54. •

We are going to prove the converse of Proposition 10.85 by showing that there is a bijection $\psi : e(C, A) \to \operatorname{Ext}^1(C, A)$. Let us construct the function ψ.

Given an extension $\xi : 0 \to A \to B \to C \to 0$ and a projective resolution of C, form the diagram

$$\begin{array}{ccccccccc}
P_2 & \xrightarrow{d_2} & P_1 & \xrightarrow{d_1} & P_0 & \longrightarrow & C & \longrightarrow & 0 \\
\vdots & & \vdots{\scriptstyle \alpha} & & \vdots & & \downarrow{\scriptstyle 1_C} & & \\
0 & \longrightarrow & A & \longrightarrow & B & \longrightarrow & C & \longrightarrow & 0
\end{array}$$

By the comparison theorem (Theorem 10.46), we may fill in dotted arrows to obtain a commutative diagram. In particular, there is a map $\alpha\colon P_1 \to A$ with $\alpha d_2 = 0$; that is, $d_2^*(\alpha) = 0$, so that $\alpha \in \ker d_2^*$ is a cocycle. The comparison theorem also says that any two fillings in of the diagram are homotopic; thus, if $\alpha'\colon P_1 \to A$ is part of a second filling in, there are maps s_0 and s_1 with $\alpha' - \alpha = 0 \cdot s_1 + s_0 d_1 = s_0 d_1$:

$$\begin{array}{ccccc} P_2 & \xrightarrow{d_2} & P_1 & \xrightarrow{d_1} & P_0 \\ & {\scriptstyle s_1}\swarrow & {\scriptstyle \alpha'}\downarrow\downarrow{\scriptstyle \alpha} & \swarrow{\scriptstyle s_0} & \\ 0 & \longrightarrow & A & \longrightarrow & B \end{array}$$

Thus, $\alpha' - \alpha \in \operatorname{im} d_1^*$, and so the homology class $\alpha + \operatorname{im} d_1^* \in \operatorname{Ext}^1(C, A)$ is well-defined. We leave as an exercise for the reader that equivalent extensions ξ and ξ' determine the same element of Ext. Thus,

$$\psi\colon e(C, A) \to \operatorname{Ext}^1(C, A),$$

given by

$$\psi([\xi]) = \alpha + \operatorname{im} d_1^*,$$

is a well-defined function. In order to prove that ψ is a bijection, we first analyze the diagram containing the map α.

Lemma 10.87. *Let* $\Xi : \ 0 \to X_1 \xrightarrow{i} X_0 \xrightarrow{\varepsilon} C \to 0$ *be an extension of a module* X_1 *by a module* C*. Given a module* A*, consider the diagram*

$$\begin{array}{ccccccccc} \Xi : 0 & \longrightarrow & X_1 & \xrightarrow{j} & X_0 & \xrightarrow{\varepsilon} & C & \longrightarrow & 0 \\ & & \downarrow{\scriptstyle \alpha} & & & & \downarrow{\scriptstyle 1_C} & & \\ & & A & & & & C & & \end{array}$$

(i) *There exists a commutative diagram with exact rows completing the given diagram:*

$$\begin{array}{ccccccccc} 0 & \longrightarrow & X_1 & \xrightarrow{j} & X_0 & \xrightarrow{\varepsilon} & C & \longrightarrow & 0. \\ & & \downarrow{\scriptstyle \alpha} & & \downarrow{\scriptstyle \beta} & & \downarrow{\scriptstyle 1_C} & & \\ 0 & \longrightarrow & A & \xrightarrow{i} & B & \xrightarrow{\eta} & C & \longrightarrow & 0 \end{array}$$

(ii) *Any two bottom rows of completed diagrams are equivalent extensions.*

Proof. (i) We define B as the pushout of j and α. Thus, if

$$S = \big\{(\alpha x_1, -j x_1) \in A \oplus X_0 : x_1 \in X_1\big\},$$

define $B = (A \oplus X_0)/S$,

$$i : a \mapsto a + S, \quad \beta : x_0 \mapsto x_0 + S, \quad \text{and} \quad \eta : (a, x_0) + S \mapsto \varepsilon x_0.$$

That η is well-defined, that the diagram commutes, and that the bottom row is exact are left for the reader to check.

(ii) Let

$$\begin{array}{ccccccccc} 0 & \longrightarrow & X_1 & \xrightarrow{j} & X_0 & \xrightarrow{\varepsilon} & C & \longrightarrow & 0. \\ & & \downarrow{\scriptstyle \alpha} & & \downarrow{\scriptstyle \beta'} & & \downarrow{\scriptstyle 1_C} & & \\ 0 & \longrightarrow & A & \xrightarrow{i'} & B' & \xrightarrow{\eta'} & C & \longrightarrow & 0 \end{array}$$

be a second completion of the diagram. Define $f : A \oplus X_0 \to B'$ by

$$f : (a, x_0) \mapsto i'a + \beta' x_0.$$

We claim that f is surjective. If $b' \in B'$, then $\eta' b' \in C$, and so there is $x_0 \in X_0$ with $\varepsilon x_0 = \eta' b'$. Commutativity gives $\eta' \beta' x_0 = \varepsilon x_0 = \eta' b'$. Hence, $b' - \beta' x_0 \in \ker \eta' = \operatorname{im} i'$, and so there is $a \in A$ with $i'a = b' - \beta' x_0$. Therefore, $b' = i'a + \beta' x_0 \in \operatorname{im} f$, as desired.

We now show that $\ker f = S$. If $(\alpha x_1, -j x_1) \in S$, then $f(\alpha x_1, -j x_1) = i'\alpha x_1 - \beta' j x_1 = 0$, by commutativity of the first square of the diagram, and so $S \subseteq \ker f$. For the reverse inclusion, let $(a, x_0) \in \ker f$, so that $i'a + \beta' x_0 = 0$. Commutativity of the second square gives $\varepsilon x_0 = \eta' \beta' x_0 = -\eta' i a = 0$. Hence, $x_0 \in \ker \varepsilon = \operatorname{im} j$, so there is $x_1 \in X_1$ with $j x_1 = x_0$. Thus, $i'a = -\beta' x_0 = -\beta' j x_1 = -i' \alpha x_1$. Since i' is injective, we have $a = -\alpha x_1$; replacing x_1 by $y_1 = -x_1$. We have $(a, x_0) = (\alpha y_1, -j y_1) \in S$, as desired.

Finally, define $\varphi : B \to B'$ by

$$\varphi : (a, x_0) + S \mapsto f(a, x_0) = i'a + \beta' x_0$$

[φ is well-defined because $B = (A \oplus X_0)/S$ and $S = \ker f$]. To show commutativity of the diagram

$$\begin{array}{ccccccccc} 0 & \longrightarrow & A & \xrightarrow{i} & B & \xrightarrow{\eta} & C & \longrightarrow & 0 \\ & & {\scriptstyle 1_A}\downarrow & & \downarrow{\scriptstyle \varphi} & & \downarrow{\scriptstyle 1_C} & & \\ 0 & \longrightarrow & A & \xrightarrow{i'} & B' & \xrightarrow{\eta'} & C & \longrightarrow & 0 \end{array}$$

we use the definitions of the maps i and η in part (i). For the first square, if $a \in A$, then $\varphi i a = \varphi((a, 0) + S) = i'a$. For the second square,

$$\begin{aligned} \eta' \varphi : (a, x_0) + S \mapsto \; & \eta'(i'a + \beta' x_0) \\ & = \eta' \beta' x_0 \\ & = \varepsilon x_0 \\ & = \eta((a, x_0) + S). \end{aligned}$$

Therefore, the two bottom rows are equivalent extensions. •

Notation. Denote the extension of A by C just constructed by

$$\alpha \Xi.$$

The dual result is true; it is related to the construction of Ext using injective resolutions of the second variable A.

Lemma 10.88. *Let A and Y_0 be modules, and let $\Xi' : 0 \to A \to Y_0 \to Y_1 \to 0$ be an extension of A by Y_1. Given a module C, consider the diagram*

$$\begin{array}{ccccccccc} & & A & & & & C & & \\ & & \Big\downarrow{\scriptstyle 1_A} & & & & \Big\downarrow{\scriptstyle \gamma} & & \\ \Xi' : 0 & \longrightarrow & A & \longrightarrow & Y_0 & \xrightarrow{p} & Y_1 & \longrightarrow & 0 \end{array}$$

(i) *There exists a commutative diagram with exact rows completing the given diagram:*

$$\begin{array}{lccccccccc} \Xi'\gamma : & 0 & \longrightarrow & A & \longrightarrow & B & \longrightarrow & C & \longrightarrow & 0 \\ & & & \Big\downarrow{\scriptstyle 1_A} & & \Big\downarrow & & \Big\downarrow{\scriptstyle \gamma} & & \\ \Xi' : & 0 & \longrightarrow & A & \longrightarrow & Y_0 & \xrightarrow{p} & Y_1 & \longrightarrow & 0 \end{array}$$

(ii) *Any two top rows of completed diagrams are equivalent extensions.*

Proof. Dual to that of Lemma 10.87; in particular, construct the top row using the pullback of γ and p. •

Notation. Denote the extension of A by C just constructed by

$$\Xi'\gamma.$$

Theorem 10.89. *The function $\psi : e(C, A) \to \operatorname{Ext}^1(C, A)$ is a bijection.*

Proof. We construct an inverse $\theta : \operatorname{Ext}^1(C, A) \to e(C, A)$ for ψ. Choose a projective resolution of C, so there is an exact sequence

$$\cdots \to P_2 \xrightarrow{d_2} P_1 \xrightarrow{d_1} P_0 \to C \to 0,$$

and choose a 1-cocycle $\alpha : P_1 \to A$. Since α is a cocycle, we have $0 = d_2^*(\alpha) = \alpha d_2$, so that α induces a homomorphism $\alpha' : P_1/\operatorname{im} d_2 \to A$ [if $x_1 \in P_1$, then $\alpha' : x_1 + \operatorname{im} d_2 \mapsto \alpha(x_1)$]. Let Ξ denote the extension

$$\Xi : 0 \to P_1/\operatorname{im} d_2 \to P_0 \to C \to 0.$$

As in the lemma, there is a commutative diagram with exact rows:

$$\begin{array}{ccccccccc} 0 & \longrightarrow & P_1/\operatorname{im} d_2 & \longrightarrow & P_0 & \longrightarrow & C & \longrightarrow & 0 \\ & & {\scriptstyle \alpha'}\downarrow & & \downarrow{\scriptstyle \beta} & & \downarrow{\scriptstyle 1_C} & & \\ 0 & \longrightarrow & A & \xrightarrow{\ i\ } & B & \longrightarrow & C & \longrightarrow & 0 \end{array}$$

Define $\theta\colon \operatorname{Ext}^1(C, A) \to e(C, A)$ using the construction in the lemma:

$$\theta(\alpha + \operatorname{im} d_1^*) = [\alpha'\Xi].$$

We begin by showing that θ is independent of the choice of cocycle α. Suppose that $\zeta\colon P_1 \to A$ is another cocycle. Now α and ζ are parts of a chain map that the comparison theorem says are homotopic. Hence, there is a map $s\colon P_0 \to A$ with $\zeta = \alpha + sd_1$. But it is easy to see that the following diagram commutes:

$$\begin{array}{ccccccccc} P_2 & \xrightarrow{d_2} & P_1 & \xrightarrow{d_1} & P_0 & \longrightarrow & C & \longrightarrow & 0 \\ \downarrow & & \downarrow{\scriptstyle \alpha+sd_1} & & \downarrow{\scriptstyle \beta+is} & & \downarrow{\scriptstyle 1_C} & & \\ 0 & \longrightarrow & A & \xrightarrow{\ i\ } & B & \longrightarrow & C & \longrightarrow & 0 \end{array}$$

As the bottom row has not changed, we have $[\alpha'\Xi] = [\zeta'\Xi]$.

It remains to show that the composites $\psi\theta$ and $\theta\psi$ are identities. If $\alpha + \operatorname{im} d_1^* \in \operatorname{Ext}^1(C, A)$, then $\theta(\alpha + \operatorname{im} d_1^*)$ is the bottom row of the diagram

$$\begin{array}{ccccccccc} 0 & \longrightarrow & P_1/\operatorname{im} d_2 & \longrightarrow & P_0 & \longrightarrow & C & \longrightarrow & 0 \\ & & {\scriptstyle \alpha'}\downarrow & & \downarrow{\scriptstyle \beta} & & \downarrow{\scriptstyle 1_C} & & \\ 0 & \longrightarrow & A & \xrightarrow{\ i\ } & B & \longrightarrow & C & \longrightarrow & 0 \end{array}$$

and $\psi\theta(\alpha + \operatorname{im} d_1^*)$ is the homology class of a cocycle fitting this diagram. Clearly, α is such a cocycle; and so $\psi\theta$ is the identity. For the other composite, start with an extension ξ, and then imbed it as the bottom row of a diagram

$$\begin{array}{ccccccccc} P_2 & \xrightarrow{d_2} & P_1 & \xrightarrow{d_1} & P_0 & \longrightarrow & C & \longrightarrow & 0 \\ \vdots & & {\scriptstyle \alpha}\vdots & & \vdots & & \downarrow{\scriptstyle 1_C} & & \\ 0 & \longrightarrow & A & \xrightarrow{\ i\ } & B & \longrightarrow & C & \longrightarrow & 0 \end{array}$$

Both ξ and $\alpha'\Xi$ are bottom rows of such a diagram, and so Lemma 10.87(ii) shows that $[\xi] = [\alpha'\Xi]$. •

We can now prove the converse of Proposition 10.85.

Corollary 10.90. *For any modules C and A, every extension of A by C is split if and only if* $\mathrm{Ext}^1_R(C, A) = \{0\}$.

Proof. Since $\mathrm{Ext}^1(C, A) = \{0\}$, we have $|e(C, A)| = 1$, so there is only one equivalence class of extensions of A by C. But the split extension always exists, and so every extension is equivalent to the split extension; that is, every extension of A by C splits. •

Example 10.91.
If p is a prime, then $\mathrm{Ext}^1_{\mathbb{Z}}(\mathbb{I}_p, \mathbb{I}_p) \cong \mathbb{I}_p$, as we saw in Example 10.84(i). On the other hand, it follows from Theorem 10.89 that there are p equivalence classes of extensions $0 \to \mathbb{I}_p \to B \to \mathbb{I}_p \to 0$. But $|B| = p^2$, so there are only two choices for B to isomorphism: $B \cong \mathbb{I}_{p^2}$ or $B \cong \mathbb{I}_p \oplus \mathbb{I}_p$. Of course, this is consistent with Example 10.18. ◂

Here is a minor application of Ext.

Proposition 10.92.

(i) *If F is a torsion-free abelian group and T is an abelian group of **bounded order** (that is, $nT = \{0\}$ for some positive integer n), then* $\mathrm{Ext}^1(F, T) = \{0\}$.

(ii) *Let G be an abelian group. If the torsion subgroup tG of G is of bounded order, then tG is a direct summand of G.*

Proof. (i) Since F is torsion-free, it is a flat $\mathbb{Z}$-module, by Corollary 9.6, so that exactness of $0 \to \mathbb{Z} \to \mathbb{Q}$ gives exactness of $0 \to \mathbb{Z} \otimes F \to \mathbb{Q} \otimes F$. Thus, $F \cong \mathbb{Z} \otimes F$ can be imbedded in a vector space V over $\mathbb{Q}$, namely, $V = \mathbb{Q} \otimes F$. Applying the contravariant functor $\mathrm{Hom}(\ , T)$ to $0 \to F \to V \to V/F \to 0$ gives an exact sequence

$$\mathrm{Ext}^1(V, T) \to \mathrm{Ext}^1(F, T) \to \mathrm{Ext}^2(V/F, T).$$

Now the last term is $\{0\}$, by Exercise 10.39 on page 852, and $\mathrm{Ext}^1(V, T)$ is (torsion-free) divisible, by Example 10.70, so that $\mathrm{Ext}^1(F, T)$ is divisible. Since T has bounded order, Exercise 10.41 on page 852 gives $\mathrm{Ext}^1(F, T) = \{0\}$.

(ii) To prove that the extension $0 \to tG \to G \to G/tG \to 0$ splits, it suffices to prove that $\mathrm{Ext}^1(G/tG, tG) = \{0\}$. Since G/tG is torsion-free, this follows from part (i) and Corollary 10.90. •

The torsion subgroup of a group may not be a direct summand; the following proof by homology is quite different from that of Exercise 9.1(iii) on page 663.

Proposition 10.93. *There exists an abelian group G whose torsion subgroup is not a direct summand of G; in fact, we may choose $tG = \sum_p \mathbb{I}_p$, where the sum is over all primes p.*

Proof. It suffices to prove that $\mathrm{Ext}^1\big(\mathbb{Q}, \sum_p \mathbb{I}_p\big) \neq 0$, for this will give a nonsplit extension $0 \to \sum_p \mathbb{I}_p \to G \to \mathbb{Q} \to 0$; moreover, since $\mathbb{Q}$ is torsion-free, it follows that $\sum_p \mathbb{I}_p = tG$.

Consider the exact sequence $0 \to \sum_p \mathbb{I}_p \to \prod_p \mathbb{I}_p \to D \to 0$. By Exercise 9.6 on page 663, we know that D is divisible (in truth, $D \cong \mathbb{R}$: it is a torsion-free divisible group, hence it is a vector space over $\mathbb{Q}$, by Proposition 9.23, and we check that $\dim(D) =$ continuum, which is the dimension of $\mathbb{R}$ as a vector space over $\mathbb{Q}$). There is an exact sequence

$$\mathrm{Hom}\big(\mathbb{Q}, \prod_p \mathbb{I}_p\big) \to \mathrm{Hom}(\mathbb{Q}, D) \xrightarrow{\partial} \mathrm{Ext}^1\big(\mathbb{Q}, \sum_p \mathbb{I}_p\big) \to \mathrm{Ext}^1(\mathbb{Q}, \prod \mathbb{I}_p).$$

But ∂ is an isomorphism: $\mathrm{Ext}^1\big(\mathbb{Q}, \prod_p \mathbb{I}_p\big) \cong \prod \mathrm{Ext}^1(\mathbb{Q}, \mathbb{I}_p) = \{0\}$, by Proposition 10.81 and Proposition 10.92, and $\mathrm{Hom}(\mathbb{Q}, \prod_p \mathbb{I}_p) \cong \prod \mathrm{Hom}(\mathbb{Q}, \mathbb{I}_p) = \{0\}$, by Theorem 7.33. Since $\mathrm{Hom}(\mathbb{Q}, D) \neq \{0\}$, we have $\mathrm{Ext}^1(\mathbb{Q}, \sum_p \mathbb{I}_p) \neq \{0\}$. •

Remark. We can prove that a torsion abelian group T has the property that it is a direct summand of any group containing it as its torsion subgroup if and only if $T \cong B \oplus D$, where B has bounded order and D is divisible. ◂

If $\mathcal{E}$ is a set and $\psi : \mathcal{E} \to G$ is a bijection to a group G, then there is a unique group structure on $\mathcal{E}$ that makes it a group and ψ an isomorphism [if $e, e' \in \mathcal{E}$, then $e = \psi^{-1}(g)$ and $e' = \psi^{-1}(g')$; define $ee' = \psi^{-1}(gg')$]. In particular, Theorem 10.89 implies that there is a group structure on $e(C, A)$; here are the necessary definitions.

Define the ***diagonal map*** $\Delta_C : C \to C \oplus C$ by $\Delta_C : c \mapsto (c, c)$, and define the ***codiagonal map*** $\nabla_A : A \oplus A \to A$ by $\nabla_A : (a_1, a_2) \mapsto a_1 + a_2$. Note that if $f, f' : C \to A$ is a homomorphism, then the composite $\nabla_A(f \oplus f')\Delta$ maps $C \to C \oplus C \to A \oplus A \to A$. It is easy to check that $\nabla_A(f \oplus f')\Delta = f + f'$, so that this formula describes addition in $\mathrm{Hom}(C, A)$. Now Ext is a generalized Hom, and so we mimic this definition to define addition in $e(C, A)$.

If $\xi : 0 \to A \to B \to C \to 0$ and $\xi' : 0 \to A' \to B' \to C' \to 0$ are extensions, then their ***direct sum*** is the extension

$$\xi \oplus \xi' : 0 \to A \oplus A' \to B \oplus B' \to C \oplus C' \to 0.$$

The ***Baer sum*** $[\xi] + [\xi']$ is defined to the equivalence class $[\nabla_A(\xi \oplus \xi')\Delta_C]$ (we have already defined $\alpha\Xi$ and $\Xi'\gamma$). To show that Baer sum is well-defined, we first show that $\alpha(\Xi'\gamma)$ is equivalent to $(\alpha\Xi')\gamma$. We then show that $e(C, A)$ is a group under this operation by showing that $\psi([\xi] + [\xi']) = \psi([\nabla_A(\xi \oplus \xi')\Delta_C])$. The identity element is the class of the split extension, and the inverse of $[\xi]$ is $[(-1_A)\xi]$.

This description of Ext^1 has been generalized by N. Yoneda to a description of Ext^n for all n. Elements of Yoneda's $\mathrm{Ext}^n(C, A)$ are certain equivalence classes of exact sequences

$$0 \to A \to B_1 \to \cdots \to B_n \to C \to 0,$$

and we add them by a generalized Baer sum (see Mac Lane, *Homology*, pages 82–87). Thus, there is a construction of Ext that does not use derived functors. Indeed, we can construct Ext^n without using projectives or injectives.

In their investigation of finite-dimensional algebras, M. Auslander and I. Reiten introduced the following notion.

Definition. An exact sequence of left R-modules, over any ring R,

$$\Xi: \quad 0 \to N \to X \to M \to 0$$

is ***almost split*** if it is not split, if both N and M are indecomposable modules, and for for all R-modules C and every R-map $\varphi\colon C \to M$ that is not an isomorphism, the exact sequence $\Xi\varphi$ is split.

Another way to say this is that $[\Xi]$ is a nonzero element of $\mathrm{Ext}^1_R(N, M)$, where N and M are indecomposable, and $[\Xi] \in \ker\varphi^*$ for every $\varphi\colon C \to M$ that is not an isomorphism. Auslander and Reiten proved that for every indecomposable module M that is not projective, there exists an almost split exact sequence ending with M. Dually, they proved that for every indecomposable module N that is not injective, there exists an almost split exact sequence beginning with N.

It is now Tor's turn. We begin with a result that has no analog for Ext.

Theorem 10.94. *If R is a ring, A is a right R-module, and B is a left R-module, then*

$$\mathrm{Tor}_n^R(A, B) \cong \mathrm{Tor}_n^{R^{\mathrm{op}}}(B, A)$$

for all $n \geq 0$, where R^{op} is the opposite ring of R.

Proof. Recall Proposition 8.11: Every left R-module is a right R^{op}-module, and every right R-module is a left R^{op}-module. Choose a deleted projective resolution $\mathbf{P}_A$ of A. It is easy to see that $t\colon \mathbf{P}_A \otimes_R B \to B \otimes_{R^{\mathrm{op}}} \mathbf{P}_A$ is a chain map of $\mathbb{Z}$-complexes, where

$$t_n\colon P_n \otimes_R B \to B \otimes_{R^{\mathrm{op}}} P_n$$

is given by

$$t_n\colon x_n \otimes b \mapsto b \otimes x_n.$$

Since each t_n is an isomorphism of abelian groups (its inverse is $b \otimes x_n \mapsto x_n \otimes b$), the chain map t is an isomorphism of complexes. By Exercise 10.22 on page 827,

$$\mathrm{Tor}_n^R(A, B) = H_n(\mathbf{P}_A \otimes_R B) \cong H_n(B \otimes_{R^{\mathrm{op}}} \mathbf{P}_A)$$

for all n. But $\mathbf{P}_A$, viewed as a complex of left R^{op}-modules, is a deleted projective resolution of A qua left R^{op}-module, and so $H_n(B \otimes_{R^{\mathrm{op}}} \mathbf{P}_A) \cong \mathrm{Tor}_n^{R^{\mathrm{op}}}(B, A)$. •

In light of this result, theorems about $\mathrm{Tor}(A, \ )$ will yield results about $\mathrm{Tor}(\ , B)$; we will not have to say "similarly in the other variable."

Corollary 10.95. *If R is a commutative ring and A and B are R-modules, then for all $n \geq 0$,*

$$\mathrm{Tor}_n^R(A, B) \cong \mathrm{Tor}_n^R(B, A).$$

We know that Tor_n vanishes on projectives; we now show that it vanishes on flat modules.

Proposition 10.96. *A right R-module F is flat if and only if $\mathrm{Tor}_n^R(A, M) = \{0\}$ for all $n \geq 1$ and every left R-module M.*

Proof. Let $0 \to N \xrightarrow{i} P \to M \to 0$ be exact, where P is projective. There is an exact sequence

$$\mathrm{Tor}_1(F, P) \to \mathrm{Tor}_1(F, M) \to F \otimes N \xrightarrow{1 \otimes i} F \otimes P.$$

Now $\mathrm{Tor}_1(F, P) = \{0\}$, because P is projective, so that $\mathrm{Tor}_1(F, M) = \ker(1 \otimes i)$. Since F is flat, however, $\ker(1 \otimes i) = \{0\}$, and so $\mathrm{Tor}_1(F, M) = \{0\}$. The result for all $n \geq 1$ follows by dimension shifting.

For the converse, $0 \to A \xrightarrow{i} B$ exact implies exactness of

$$0 = \mathrm{Tor}_1(F, B/A) \to F \otimes A \xrightarrow{1 \otimes i} F \otimes B.$$

Hence, $1 \otimes i$ is an injection, and so F is flat. (Notice that we have only assumed the vanishing of Tor_1 in proving the converse.) •

Proposition 10.97. *If $\{B_k : k \in K\}$ is a family of left R-modules, then there are natural isomorphisms, for all n,*

$$\mathrm{Tor}_n^R\big(A, \sum_{k \in K} B_k\big) \cong \sum_{k \in K} \mathrm{Tor}_n^R(A, B_k).$$

There is also an isomorphism if the sum is in the first variable.

Proof. The proof is by dimension shifting. The base step is Theorem 8.87, for $\mathrm{Tor}_0(A, \ )$ is naturally equivalent to $A \otimes \ $.

For the inductive step, choose, for each $k \in K$, a short exact sequence

$$0 \to N_k \to P_k \to B_k \to 0,$$

where P_k is projective. There is an exact sequence

$$0 \to \sum_k N_k \to \sum_k P_k \to \sum_k B_k \to 0,$$

and $\sum_k P_k$ is projective, for every sum of projectives is projective. There is a commutative diagram with exact rows:

$$\begin{array}{ccccccc}
\mathrm{Tor}_1(A, \sum P_k) & \longrightarrow & \mathrm{Tor}_1(A, \sum B_k) & \xrightarrow{\partial} & A \otimes \sum N_k & \longrightarrow & A \otimes \sum P_k \\
 & & \vdots & & \downarrow \tau & & \downarrow \sigma \\
\sum \mathrm{Tor}_1(A, P_k) & \longrightarrow & \sum \mathrm{Tor}_1(A, B_k) & \xrightarrow{\partial'} & \sum A \otimes N_k & \longrightarrow & \sum A \otimes P_k,
\end{array}$$

where the maps in the bottom row are just the usual induced maps in each coordinate, and the maps τ and σ are the isomorphisms given by Theorem 8.87. The proof is completed by dimension shifting. •

Example 10.98.

(i) We show, for every abelian group B, that

$$\mathrm{Tor}_1^{\mathbb{Z}}(\mathbb{I}_n, B) \cong B[n] = \{b \in B : nb = 0\}.$$

There is an exact sequence

$$0 \to \mathbb{Z} \xrightarrow{\mu_n} \mathbb{Z} \to \mathbb{I}_n \to 0,$$

where μ_n is multiplication by n. Applying $\otimes B$ gives exactness of

$$\mathrm{Tor}_1(\mathbb{Z}, B) \to \mathrm{Tor}_1(\mathbb{I}_n, B) \to \mathbb{Z} \otimes B \xrightarrow{1 \otimes \mu_n} \mathbb{Z} \otimes B.$$

Now $\mathrm{Tor}_1(\mathbb{Z}, B) = \{0\}$, because $\mathbb{Z}$ is projective. Moreover, $1 \otimes \mu_n$ is also multiplication by n, while $\mathbb{Z} \otimes B = B$. More precisely, $\mathbb{Z}\otimes$ is naturally equivalent to the identity functor on **Ab**, and so there is a commutative diagram with exact rows

$$\begin{array}{ccccccc}
0 & \longrightarrow & B[n] & \longrightarrow & B & \xrightarrow{\mu_n} & B \\
 & & \vdots & & \downarrow{\scriptstyle \tau_B} & & \downarrow{\scriptstyle \tau_B} \\
0 & \longrightarrow & \mathrm{Tor}_1(\mathbb{I}_n, B) & \longrightarrow & \mathbb{Z} \otimes B & \xrightarrow{1 \otimes \mu_n} & \mathbb{Z} \otimes B
\end{array}$$

By Proposition 8.94, there is an isomorphism $B[n] \cong \mathrm{Tor}_1(\mathbb{I}_n, B)$.

(ii) We can now compute $\mathrm{Tor}_1^{\mathbb{Z}}(A, B)$ whenever A and B are finitely generated abelian groups. By the fundamental theorem, both A and B are direct sums of cyclic groups. Since Tor commutes with direct sums, $\mathrm{Tor}_1^{\mathbb{Z}}(A, B)$ is the direct sum of groups $\mathrm{Tor}_1^{\mathbb{Z}}(C, D)$, where C and D are cyclic. We may assume that C and D are finite, otherwise they are projective and $\mathrm{Tor}_1 = \{0\}$. This calculation can be completed using part (i) and Exercise 5.6 on page 267, which says that if D is a cyclic group of finite order m, then $D[n]$ is a cyclic group of order d, where $d = (m, n)$ is their gcd. ◂

In contrast to Ext, Proposition 10.97 can be generalized by replacing sums by direct limits.

Proposition 10.99. *If $\{B_i, \varphi_j^i\}$ is a direct system of left R-modules over a directed index set I, then there is an isomorphism, for all right R-modules A and for all $n \geq 0$,*

$$\mathrm{Tor}_n^R\big(A, \varinjlim B_i\big) \cong \varinjlim \mathrm{Tor}_n^R(A, B_i).$$

Proof. The proof is by dimension shifting. The base step is Theorem 8.101, for $\mathrm{Tor}_0(A, \ )$ is naturally equivalent to $A\otimes$.

For the inductive step, choose, for each $i \in I$, a short exact sequence

$$0 \to N_i \to P_i \to B_i \to 0,$$

where P_i is projective. Since the index set is directed, Proposition 7.100 says that there is an exact sequence

$$0 \to \varinjlim N_i \to \varinjlim P_i \to \varinjlim B_i \to 0.$$

Now $\varinjlim P_i$ is flat, for every projective module is flat, and a direct limit of flat modules is flat, by Corollary 8.102. There is a commutative diagram with exact rows:

$$\begin{array}{ccccccc} \mathrm{Tor}_1(A, \varinjlim P_i) & \longrightarrow & \mathrm{Tor}_1(A, \varinjlim B_i) & \xrightarrow{\partial} & A \otimes \varinjlim N_i & \longrightarrow & A \otimes \varinjlim P_i \\ & & \Big\downarrow & & \Big\downarrow \tau & & \Big\downarrow \sigma \\ \varinjlim \mathrm{Tor}_1(A, P_i) & \longrightarrow & \varinjlim \mathrm{Tor}_1(A, B_i) & \xrightarrow{\vec{\partial}} & \varinjlim A \otimes N_i & \longrightarrow & \varinjlim A \otimes P_i, \end{array}$$

where the maps in the bottom row are just the usual induced maps between direct limits, and the maps τ and σ are the isomorphisms given by Theorem 8.101. The step $n \geq 2$ is routine. •

This last proposition generalizes Lemma 8.97, which says that if every finitely generated submodule of a module M is flat, then M itself is flat. After all, by Example 7.97(iv), M is a direct limit, over a directed index set, of its finitely generated submodules.

When the ring R is noncommutative, $A \otimes_R B$ is an abelian group, but it need not be an R-module.

Proposition 10.100.

(i) *Let $r \in Z(R)$ be a central element, let A be a right R-module, and let B be a left R-modules. If $\mu_r : B \to B$ is multiplication by r, then the induced map*

$$\mu_{r*} \colon \mathrm{Tor}_n^R(A, B) \to \mathrm{Tor}_n^R(A, B)$$

is also multiplication by r.

(ii) *If R is a commutative ring, then $\mathrm{Tor}_n^R(A, B)$ is an R-module.*

Proof. (i) This follows at once from Example 10.47.

(ii) This follows from part (i) if we define scalar multiplication by r to be μ_{r*}. •

We are now going to assume that R is a domain, so that the notion of torsion submodule is defined, and we shall see why Tor is so called.

Lemma 10.101. *Let R be a domain, let $Q = \text{Frac}(R)$, and let $K = Q/R$.*

(i) *If A is a torsion R-module, then $\text{Tor}_1^R(K, A) \cong A$.*

(ii) *For every R-module A, we have $\text{Tor}_n(K, A) = \{0\}$ for all $n \geq 2$.*

(iii) *If A is a torsion-free R-module, then $\text{Tor}_1(K, A) = \{0\}$.*

Proof. (i) Exactness of $0 \to R \to Q \to K \to 0$ gives exactness of

$$\text{Tor}_1(Q, A) \to \text{Tor}_1(K, A) \to R \otimes A \to Q \otimes A.$$

Now Q is flat, by Corollary 8.103, and so $\text{Tor}_1(Q, A) = \{0\}$, by Proposition 10.96. The last term $Q \otimes A = \{0\}$ because Q is divisible and A is torsion, by Exercise 9.15 on page 665, and so the middle map $\text{Tor}_1(K, A) \to R \otimes A$ is an isomorphism.

(ii) There is an exact sequence

$$\text{Tor}_n(Q, A) \to \text{Tor}_n(K, A) \to \text{Tor}_{n-1}(R, A).$$

Since $n \geq 2$, we have $n - 1 \geq 1$, and so both the first and third Tor's are $\{0\}$, because Q and R are flat. Therefore, exactness gives $\text{Tor}_n(K, A) = \{0\}$.

(iii) By Theorem 8.104, there is an injective R-module E containing A as a submodule. Since A is torsion-free, however, $A \cap tE = \{0\}$, and so A is imbedded in E/tE. By Lemma 7.72, injective modules are divisible, and so E is divisible, as is its quotient E/tE. Now E/tE is a vector space over Q, for it is a torsion-free divisible R-module (Exercise 9.7 on page 664). Let us denote E/tE by V. Since every vector space has a basis, V is a direct sum of copies of Q. Corollary 8.103 says that Q is flat, and Lemma 8.98 says that a direct sum of flat modules is flat. We conclude that V is flat.[14]

Exactness of $0 \to A \to V \to V/A \to 0$ gives exactness of

$$\text{Tor}_2(K, V/A) \to \text{Tor}_1(K, A) \to \text{Tor}_1(K, V).$$

Now $\text{Tor}_2(K, V/A) = \{0\}$, by part (ii), and $\text{Tor}_1(K, V) = \{0\}$, because V is flat. We conclude from exactness that $\text{Tor}_1(K, A) = \{0\}$. •

The next result shows why Tor is so-called.

Theorem 10.102.

(i) *If R is a domain, $Q = \text{Frac}(R)$, and $K = Q/R$, then the functor $\text{Tor}_1^R(K, \ )$ is naturally equivalent to the torsion functor.*

(ii) *$\text{Tor}_1^R(K, A) \cong tA$ for all R-modules A.*

[14]Torsion-free $\mathbb{Z}$-modules are flat, but there exist domains R having torsion-free modules that are not flat. In fact, domains for which every torsion-free module is flat, called *Prüfer rings*, are characterized as those domains in which every finitely generated ideal is a projective module.

Proof. Exactness of

$$\mathrm{Tor}_2(K, A/tA) \to \mathrm{Tor}_1(K, tA) \xrightarrow{\iota_A} \mathrm{Tor}_1(K, A) \to \mathrm{Tor}_1(K, A/tA).$$

The first term is {0}, by Lemma 10.101(ii), and the last term is {0}, by Lemma 10.101(iii). Therefore, the map $\iota_A\colon \mathrm{Tor}_1(K, tA) \to \mathrm{Tor}_1(K, A)$ is an isomorphism.

Let $f\colon A \to B$ and let $f'\colon tA \to tB$ be its restriction. The following diagram commutes, because $\mathrm{Tor}_1(K, \ )$ is a functor, and this says that the isomorphisms ι_A constitute a natural transformation.

$$\begin{array}{ccc} \mathrm{Tor}_1(K, tA) & \xrightarrow{\iota_A} & \mathrm{Tor}_1(K, A) \\ {\scriptstyle f'_*}\downarrow & & \downarrow{\scriptstyle f_*} \\ \mathrm{Tor}_1(K, tB) & \xrightarrow{\iota_B} & \mathrm{Tor}_1(K, B) \end{array}$$ •

There is a construction of $\mathrm{Tor}_1^{\mathbb{Z}}(A, B)$ by generators and relations. Consider all triples (a, n, b), where $a \in A$, $b \in B$, $na = 0$, and $nb = 0$. Then $\mathrm{Tor}_1^{\mathbb{Z}}(A, B)$ is generated by all such triples subject to the relations (whenever both sides are defined):

$$(a + a', n, b) = (a, n, b) + (a', n, b)$$
$$(a, n, b + b') = (a, n, b) + (a, n, b')$$
$$(ma, n, b) = (a, mn, b) = (a, m, nb).$$

For a proof of this result, and its generalization to $\mathrm{Tor}_n^R(A, B)$ for arbitrary rings R, see Mac Lane, *Homology*, pp. 150–159.

The Tor functors are very useful in algebraic topology. The *Universal Coefficients Theorem* gives a formula for the homology groups $H_n(X; G)$ with coefficients in an abelian group G.

Theorem (Universal Coefficients). *For every topological space X and every abelian group G, there are isomorphisms for all $n \geq 0$,*

$$H_n(X; G) \cong H_n(X) \otimes_{\mathbb{Z}} G \,\oplus\, \mathrm{Tor}_1^{\mathbb{Z}}(H_{n-1}(X), G).$$

Proof. See Rotman, *An Introduction to Algebraic Topology*, page 261. •

If we know the homology groups of spaces X and Y, then the *Künneth formula* gives a formula for the homology groups of $X \times Y$, and this, too, involves Tor in an essential way.

Theorem (Künneth Formula). *For every pair of topological spaces X and Y, there are isomorphisms for every $n \geq 0$,*

$$H_n(X \times Y) \cong \sum_i H_i(X) \otimes_{\mathbb{Z}} H_{n-i}(Y) \,\oplus\, \sum_p \mathrm{Tor}_1^{\mathbb{Z}}(H_p(X), H_{n-1-p}(Y)).$$

Proof. See Rotman, *An Introduction to Algebraic Topology*, page 269 •

EXERCISES

10.44 Prove the following analog of Theorem 10.45. Let $\mathcal{E}^n\colon {}_R\mathbf{Mod} \to \mathbf{Ab}$ be a sequence of covariant functors, for $n \geq 0$, such that

(i) for every short exact sequence $0 \to A \to B \to C \to 0$, there is a long exact sequence and natural connecting homomorphisms

$$\cdots \to \mathcal{E}^n(A) \to \mathcal{E}^n(B) \to \mathcal{E}^n(C) \xrightarrow{\Delta_n} \mathcal{E}^{n+1}(A) \to \cdots;$$

(ii) there is a left R-module M such that $\mathcal{E}^0$ and $\mathrm{Hom}_R(M,\ )$ are naturally equivalent;

(iii) $\mathcal{E}^n(E) = \{0\}$ for all injective modules E and all $n \geq 1$.

Prove that $\mathcal{E}^n$ is naturally equivalent to $\mathrm{Ext}^n(M,\ )$ for all $n \geq 0$.

10.45 Let $\mathrm{TOR}^n\colon {}_R\mathbf{Mod} \to \mathbf{Ab}$ be a sequence of covariant functors, for $n \geq 0$, such that

(i) for every short exact sequence $0 \to A \to B \to C \to 0$, there is a long exact sequence and natural connecting homomorphisms

$$\cdots \to \mathrm{TOR}_n(A) \to \mathrm{TOR}_n(B) \to \mathrm{TOR}_n(C) \xrightarrow{\Delta_n} \mathrm{TOR}_{n-1}(A) \to \cdots;$$

(ii) there is a left R-module M such that TOR_0 and $\ \otimes_R M$ are naturally equivalent;

(iii) $\mathrm{TOR}_n(P) = \{0\}$ for all projective modules P and all $n \geq 1$.

Prove that TOR_n is naturally equivalent to $\mathrm{Tor}_n(\ , M)$ for all $n \geq 0$. (There is a similar result if the first variable is fixed.)

10.46 Prove that any two split extensions of modules A by C are equivalent.

10.47 Prove that if A is an abelian group with $nA = A$ for some positive integer n, then every extension $0 \to A \to E \to \mathbb{I}_n \to 0$ splits.

10.48 If A is a torsion abelian group, prove that $\mathrm{Ext}^1(A, \mathbb{Z}) \cong \mathrm{Hom}(A, S^1)$, where S^1 is the circle group.

10.49 Prove that a left R-module E is injective if and only if $\mathrm{Ext}^1_R(A, E) = \{0\}$ for every left R-module A.

10.50 For any ring R, prove that a left R-module B is injective if and only if $\mathrm{Ext}^1(R/I, B) = \{0\}$ for every left ideal I.
Hint. Use the Baer criterion.

10.51 Prove that an abelian group G is injective if and only if $\mathrm{Ext}^1(\mathbb{Q}/\mathbb{Z}, G) = \{0\}$.

10.52 Prove that an abelian group G is free abelian if and only if $\mathrm{Ext}^1(G, F) = \{0\}$ for every free abelian group F.[15]

10.53 If $0 \to A \to B \to C \to 0$ is an exact sequence of right R-modules with both A and C flat, prove that B is flat.

10.54 If A and B are finite abelian groups, prove that $\mathrm{Tor}^{\mathbb{Z}}_1(A, B) \cong A \otimes_{\mathbb{Z}} B$.

[15]The question whether $\mathrm{Ext}^1(G, \mathbb{Z}) = \{0\}$ implies G is free abelian is known as *Whitehead's problem*. It turns out that if G countable, then it must be free abelian, but S. Shelah proved that it is undecideable whether uncountable such G must be free abelian.

10.55 Let R be a domain, $Q = \text{Frac}(R)$, and $K = Q/R$.

(i) Prove, for every R-module A, that there is an exact sequence

$$0 \to tA \to A \to Q \otimes A \to K \otimes A \to 0.$$

(ii) Prove that a module A is torsion if and only if $Q \otimes A = \{0\}$.

10.56 Let R be a domain.

(i) If B is a torsion R-module, prove that $\text{Tor}_n(A, B)$ is a torsion R-module for all R-modules A and for all $n \geq 0$.

(ii) For all R-modules A and B, prove that $\text{Tor}_n(A, B)$ is a torsion R-module for all $n \geq 1$.

10.57 Let k be a field, let $R = k[x, y]$, and let I be the ideal (x, y).

(i) Prove that $x \otimes y - y \otimes x \in I \otimes_R I$ is nonzero.
Hint. Consider $(I/I^2) \otimes (I/I^2)$.

(ii) Prove that $x(x \otimes y - y \otimes x) = 0$, and conclude that $I \otimes_R I$ is not torsion-free.

10.7 Cohomology of Groups

Recall that Proposition 10.30 and Proposition 10.31 say that there is an exact sequence

$$F_3 \xrightarrow{d_3} F_2 \xrightarrow{d_2} F_1 \xrightarrow{d_1} F_0 \to \mathbb{Z} \to 0,$$

where F_0, F_1, F_2, and F_3 are free Q-modules and $\mathbb{Z}$ is viewed as a trivial Q-module. In light of the calculations in Section 10.3, the following definition should now seem reasonable.

Definition. Let G be a group, let A be a G-module (i.e, a left $\mathbb{Z}G$-module), and let $\mathbb{Z}$ be the integers viewed as a trivial G-module (i.e, $gm = m$ for all $g \in G$ and $m \in \mathbb{Z}$). The ***cohomology groups*** of G are

$$H^n(G, A) = \text{Ext}^n_{\mathbb{Z}G}(\mathbb{Z}, A);$$

the ***homology groups*** of G are

$$H_n(G, A) = \text{Tor}^{\mathbb{Z}G}_n(\mathbb{Z}, A).$$

The history of cohomology of groups is quite interesting. The subject began with the discovery, by the topologist W. Hurewicz in the 1930's, that if X is a connected *aspherical space* (in modern language, if the higher homotopy groups of X are all trivial), then all the homology and cohomology groups of X are determined by the fundamental group $\pi = \pi_1(X)$. This led to the question of whether $H_n(X)$ could be described algebraically in terms of π. For example, Hurewicz proved that $H_1(X) \cong \pi/\pi'$, where π' is the commutator subgroup. In 1942, H. Hopf proved that if π has a presentation F/R, where F is free, then $H_2(X) \cong (R \cap F')/[F, R]$, where $[F, R]$ is the subgroup generated by all commutators of

the form $frf^{-1}r^{-1}$ for $f \in F$ and $r \in R$. These results led S. Eilenberg, S. Mac Lane, Hopf, H. Freudenthal, and B. Eckmann to create cohomology of groups.

In what follows, we will write Hom_G instead of $\operatorname{Hom}_{\mathbb{Z}G}$ and $\otimes_G$ instead of $\otimes_{\mathbb{Z}G}$. Because of the special role of the trivial G-module $\mathbb{Z}$, the augmentation

$$\varepsilon \colon \mathbb{Z}G \to \mathbb{Z},$$

defined by

$$\varepsilon \colon \sum_{x \in G} m_x x \mapsto \sum_{x \in G} m_x,$$

is important. Recall that we have seen, in Exercise 8.37 on page 573, that ε is a surjective ring homomorphism, and so its kernel, $\mathcal{G}$, is a two-sided ideal in $\mathbb{Z}G$, called the ***augmentation ideal***. Thus, there is an exact sequence

$$0 \to \mathcal{G} \to \mathbb{Z}G \xrightarrow{\varepsilon} \mathbb{Z} \to 0.$$

Proposition 10.103. *Let G be a group with augmentation ideal $\mathcal{G}$. As an abelian group, $\mathcal{G}$ is free abelian with basis $G - 1 = \{x - 1 : x \in G, x \neq 1\}$.*

Proof. An element $u = \sum_x m_x x \in \mathbb{Z}G$ lies in $\ker \varepsilon = \mathcal{G}$ if and only if $\sum_x m_x = 0$. Therefore, if $u \in \mathcal{G}$, then

$$u = u - \Big(\sum_x m_x\Big)1 = \sum_x m_x(x - 1).$$

Thus, $\mathcal{G}$ is generated by the nonzero $x - 1$ for $x \in G$.

Suppose that $\sum_{x \neq 1} n_x(x - 1) = 0$. Then $\sum_{x \neq 1} n_x x - \big(\sum_{x \neq 1} n_x\big)1 = 0$ in $\mathbb{Z}G$, which, as an abelian group, is free abelian with basis the elements of G. Hence, $n_x = 0$ for all $x \neq 1$. Therefore, the nonzero $x - 1$ comprise a basis of $\mathcal{G}$. •

We begin by examining homology groups.

Proposition 10.104. *If A is a G-module, then*

$$H_0(G, A) = \mathbb{Z} \otimes_G A \cong A/\mathcal{G}A.$$

.

Proof. By definition, $H_0(G, A) = \operatorname{Tor}_0^{\mathbb{Z}G}(\mathbb{Z}, A) = \mathbb{Z} \otimes_G A$. Applying the right exact functor $\otimes_G A$ to the exact sequence

$$0 \to \mathcal{G} \to \mathbb{Z}G \to \mathbb{Z} \to 0$$

gives exactness of the first row of the following commutative diagram:

$$\begin{array}{ccccccc}
\mathcal{G} \otimes_G A & \longrightarrow & \mathbb{Z}G \otimes_G A & \longrightarrow & \mathbb{Z} \otimes_G A & \longrightarrow & 0 \\
\downarrow & & \downarrow & & \vdots & & \\
\mathcal{G}A & \longrightarrow & A & \longrightarrow & A/\mathcal{G}A & \longrightarrow & 0
\end{array}$$

The two solid vertical arrows are given by $u \otimes a \mapsto ua$. By Proposition 8.93, there is an isomorphism $\mathbb{Z} \otimes_G A \cong A/\mathcal{G}A$. •

It is easy to see that $A/\mathcal{G}A$ is G-trivial; indeed, it is the largest G-trivial quotient of A.

Example 10.105.
Suppose that E is a semidirect product of an abelian group A by a group G. Recall that $[G, A]$ is the subgroup generated by all commutators of the form $[x, a] = xax^{-1}a^{-1}$, where $x \in G$ and $a \in A$. If we write commutators additively, as we did at the beginning of this chapter, then

$$[x, a] = x + a - x - a = xa - a = (x - 1)a$$

(recall that G acts on A by conjugation). Therefore, $A/\mathcal{G}A = A/[G, A]$ here. ◂

We are now going to use the independence of the choice of projective resolution to compute the homology groups of a finite cyclic group G.

Lemma 10.106. *Let $G = \langle x \rangle$ be a cyclic group of finite order k. Define elements D and N in $\mathbb{Z}G$ by*

$$D = x - 1 \quad \textit{and} \quad N = 1 + x + x^2 + \cdots + x^{k-1}.$$

Then the following sequence is a G-free resolution of Z:

$$\cdots \to \mathbb{Z}G \xrightarrow{N} \mathbb{Z}G \xrightarrow{D} \mathbb{Z}G \xrightarrow{N} \mathbb{Z}G \xrightarrow{D} \mathbb{Z}G \xrightarrow{\varepsilon} \mathbb{Z} \to 0,$$

where ε is the augmentation and the other maps are multiplication by N and D, respectively.

Proof. Obviously, every term $\mathbb{Z}G$ is free; moreover, since $\mathbb{Z}G$ is commutative, the maps are G-maps. Now $DN = ND = x^k - 1 = 0$, while if $u \in \mathbb{Z}G$, then

$$\varepsilon D(u) = \varepsilon\big((x - 1)u\big) = \varepsilon(x - 1)\varepsilon(u) = 0,$$

because ε is a ring map. Thus, we have a complex, and it only remains to prove exactness.

We have already noted that ε is surjective. Now $\ker \varepsilon = \mathcal{G} = \operatorname{im} D$, by Proposition 10.103, and so we have exactness at the zeroth step.

Suppose $u = \sum_{i=0}^{k-1} m_i x^i \in \ker D$; that is, $(x - 1)u = 0$. Expanding, and using the fact that $\mathbb{Z}G$ has basis $\{1, x, x^2, \ldots, x^{k-1}\}$, we have

$$m_0 = m_1 = \cdots = m_{k-1},$$

so that $u = m_0 N \in \operatorname{im} N$, as desired.

Finally, if $u = \sum_{i=0}^{k-1} m_i x^i \in \ker N$, then $0 = \varepsilon(Nu) = \varepsilon(N)\varepsilon(u) = k\varepsilon(u)$, so that $\varepsilon(u) = \sum_{i=0}^{k-1} m_i = 0$. Therefore,

$$u = -D\big(m_0 1 + (m_0 + m_1)x + \cdots + (m_0 + \cdots + m_{k-1})x^{k-1}\big) \in \operatorname{im} D. \quad \bullet$$

Definition. If A is a G-module, define submodules

$$A[N] = \{a \in A : Na = 0\}$$

and

$$A^G = \{a \in A : ga = a \text{ for all } g \in G\}.$$

Theorem 10.107. *If G is a cyclic group of finite order k and A is a G-module, then*

$$\begin{aligned} H_0(G, A) &= A/\mathcal{G}A; \\ H_{2n-1}(G, A) &= A^G/NA \quad \text{for all } n \geq 1; \\ H_{2n}(G, A) &= A[N]/\mathcal{G}A \quad \text{for all } n \geq 1. \end{aligned}$$

Proof. Apply $\otimes_G A$ to the resolution of $\mathbb{Z}$ in Lemma 10.106, noting that $\mathbb{Z}G \otimes_G A \cong A$. The calculation of ker / im is now simple, using $\operatorname{im} D = \mathcal{G}A$, which follows from Proposition 10.103 and the fact that $(x - 1) \mid (x^i - 1)$. •

Corollary 10.108. *If G is a finite cyclic group of order k and A is a trivial G-module, then*

$$\begin{aligned} H_0(G, A) &= A; \\ H_{2n-1}(G, A) &= A/kA \quad \text{for all } n \geq 1; \\ H_{2n}(G, A) &= A[k] \quad \text{for all } n \geq 1. \end{aligned}$$

In particular,

$$\begin{aligned} H_0(G, \mathbb{Z}) &= \mathbb{Z}; \\ H_{2n-1}(G, \mathbb{Z}) &= \mathbb{Z}/k\mathbb{Z} \quad \text{for all } n \geq 1; \\ H_{2n}(G, \mathbb{Z}) &= \{0\} \quad \text{for all } n \geq 1. \end{aligned}$$

Proof. Since A is G-trivial, we have $A^G = A$ and $\mathcal{G}A = \{0\}$ (for $Da = (x - 1)a = 0$ because $xa = a$). •

We now compute low-dimensional homology groups of not necessarily cyclic groups.

Lemma 10.109. *For any group G, we have*

$$H_1(G, \mathbb{Z}) \cong \mathcal{G}/\mathcal{G}^2.$$

Proof. The long exact sequence arising from

$$0 \to \mathcal{G} \to \mathbb{Z}G \xrightarrow{\varepsilon} \mathbb{Z} \to 0$$

ends with

$$H_1(G, \mathbb{Z}G) \to H_1(G, \mathbb{Z}) \xrightarrow{\partial} H_0(G, \mathcal{G}) \to H_0(G, \mathbb{Z}G) \xrightarrow{\varepsilon_*} H_0(G, \mathbb{Z}) \to 0.$$

Now $H_1(G, \mathbb{Z}G) = \{0\}$, because $\mathbb{Z}G$ is projective, so that ∂ is an injection. Also,

$$H_0(G, \mathbb{Z}G) \cong \mathbb{Z},$$

by Proposition 10.104. Since ε_* is surjective, it must be injective as well (if $\ker \varepsilon_* \neq \{0\}$, then $\mathbb{Z}/\ker \varepsilon_*$ is finite; on the other hand, $\mathbb{Z}/\ker \varepsilon_* \cong \operatorname{im} \varepsilon_* = \mathbb{Z}$, which is torsion-free). Exactness of the sequence of homology groups (∂ is surjective if and only if ε_* is injective) now gives ∂ a surjection. We conclude that

$$\partial\colon H_1(G, \mathbb{Z}) \cong H_0(G, \mathcal{G}) \cong \mathcal{G}/\mathcal{G}^2,$$

by Proposition 10.104. •

Proposition 10.110. *For any group G, we have*

$$H_1(G, \mathbb{Z}) \cong G/G',$$

where G' is the commutator subgroup of G.

Proof. It suffices to prove that $G/G' \cong \mathcal{G}/\mathcal{G}^2$. Define $\theta\colon G \to \mathcal{G}/\mathcal{G}^2$ by

$$\theta\colon x \mapsto (x - 1) + \mathcal{G}^2.$$

To see that θ is a homomorphism, note that

$$xy - 1 - (x - 1) - (y - 1) = (x - 1)(y - 1) \in \mathcal{G}^2,$$

so that

$$\begin{aligned}
\theta(xy) &= xy - 1 + \mathcal{G}^2 \\
&= (x - 1) + (y - 1) + \mathcal{G}^2 \\
&= x - 1 + \mathcal{G}^2 + y - 1 + \mathcal{G}^2 \\
&= \theta(x) + \theta(y).
\end{aligned}$$

Since $\mathcal{G}/\mathcal{G}^2$ is abelian, $\ker \theta \subseteq G'$, and so θ induces a homomorphism $\theta'\colon G/G' \to \mathcal{G}/\mathcal{G}^2$, namely, $xG' \mapsto x - 1 + \mathcal{G}^2$.

We now construct the inverse of θ'. By Proposition 10.103, $\mathcal{G}$ is a free abelian group with basis all $x - 1$, where $x \in G$ and $x \neq 1$. It follows that there is a (well-defined) homomorphism $\varphi\colon \mathcal{G} \to G/G'$, given by

$$\varphi\colon x - 1 \mapsto xG'.$$

If $\mathcal{G}^2 \subseteq \ker \varphi$, then φ induces a homomorphism $\mathcal{G}/\mathcal{G}^2 \to G/G'$ that, obviously, is the inverse of θ', and this will complete the proof.

If $u \in \mathcal{G}^2$, then

$$\begin{aligned} u &= \Big(\sum_{x \neq 1} m_x(x-1)\Big)\Big(\sum_{y \neq 1} n_y(y-1)\Big) \\ &= \sum_{x,y} m_x n_y (x-1)(y-1) \\ &= \sum_{x,y} m_x n_y \big((xy-1)-(x-1)-(y-1)\big). \end{aligned}$$

Therefore, $\varphi(u) = \prod_{x,y}(xyx^{-1}y^{-1})^{m_x n_y} G' = G'$, and so $u \in \ker \varphi$, as desired. •

The group $H_2(G, \mathbb{Z})$ is useful; it is called the ***Schur multiplier*** of G. For example, suppose that $G = F/R$, where F is a free group; that is, we have a presentation of a group G. Then ***Hopf's formula*** is

$$H_2(G, \mathbb{Z}) \cong (R \cap F)/[F, R]$$

(see Rotman, *An Introduction to Homological Algebra*, page 274). It follows that the group $(R \cap F)/[F, R]$ depends only on G and not upon the choice of presentation of G.

Definition. An exact sequence $0 \to A \to E \to G \to 1$ is a ***central extension*** of a group G if $A \leq Z(E)$. A ***universal central extension*** of G is a central extension $0 \to M \to U \to G \to 1$ for which there always exists a commutative diagram

$$\begin{array}{ccccccccc} 0 & \longrightarrow & A & \longrightarrow & E & \longrightarrow & G & \longrightarrow & 1 \\ & & \vdots & & \vdots & & \downarrow{\scriptstyle 1_G} & & \\ 0 & \longrightarrow & M & \longrightarrow & U & \longrightarrow & G & \longrightarrow & 1 \end{array}$$

Theorem. *If G is a finite group, then G has a universal central extension if and only if $G = G'$, in which case $M \cong H_2(G, \mathbb{Z})$. In particular, every finite simple group has a universal central extension.*

Proof. See Milnor, *Introduction to Algebraic K-Theory*, pages 43–46. •

This theorem is used to construct "covers" of simple groups.

We now consider cohomology groups.

Proposition 10.111. *Let G be a group, let A be a G-module, and let $\mathbb{Z}$ be viewed as a trivial G-module. Then*

$$H^0(G, A) = \operatorname{Hom}_G(\mathbb{Z}, A) \cong A^G.$$

Proof. By definition,

$$H^0(G, A) = \operatorname{Ext}^0_{\mathbb{Z}G}(\mathbb{Z}, A) = \operatorname{Hom}_G(\mathbb{Z}, A).$$

Define $\tau_A \colon \operatorname{Hom}_G(\mathbb{Z}, A) \to A^G$ by $f \mapsto f(1)$. Note that $f(1) \in A^G$: If $g \in G$, then $gf(1) = f(g \cdot 1)$ (because f is a G-map), and $g \cdot 1 = 1$ (because $\mathbb{Z}$ is G-trivial); therefore, $gf(1) = f(1)$, and $f(1) \in A^G$. That τ_A is an isomorphism is a routine calculation. •

It follows that $H^0(G, A)$ is the largest G-trivial submodule of A.

Theorem 10.112. *Let $G = \langle\sigma\rangle$ be a cyclic group of finite order k, and let A be a G-module. If $N = \sum_{i=0}^{k-1} \sigma^i$ and $D = \sigma - 1$, then*

$$\begin{aligned} H^0(G, A) &= A^G; \\ H^{2n-1}(G, A) &= \ker N/(\sigma - 1)A \quad for\ all\ n \geq 1; \\ H^{2n}(G, A) &= A^G/NA \quad for\ all\ n \geq 1. \end{aligned}$$

Proof. Apply the contravariant $\text{Hom}_G(\ , A)$ to the resolution of $\mathbb{Z}$ in Lemma 10.106, noting that $\text{Hom}_G(\mathbb{Z}G, A) \cong A$. The calculation of ker / im is now as given in the statement. •

Note that Proposition 10.103 gives $\text{im}\, D = \mathcal{G}A$.

Corollary 10.113. *If G is a cyclic group of finite order k and A is a trivial G-module, then*

$$\begin{aligned} H^0(G, A) &= A; \\ H^{2n-1}(G, A) &= A[k] \quad for\ all\ n \geq 1; \\ H^{2n}(G, A) &= A/kA \quad for\ all\ n \geq 1. \end{aligned}$$

In particular,

$$\begin{aligned} H^0(G, \mathbb{Z}) &= \mathbb{Z}; \\ H^{2n-1}(G, \mathbb{Z}) &= \{0\} \quad \textit{for all } n \geq 1; \\ H^{2n}(G, \mathbb{Z}) &= \mathbb{Z}/k\mathbb{Z} \quad \textit{for all } n \geq 1. \end{aligned}$$

Remark. A finite group G for which there exists a nonzero integer d such that

$$H^n(G, A) \cong H^{n+d}(G, A),$$

for all $n \geq 1$ and all G-modules A, is said to have ***periodic cohomology***. It can be proved that a group G has periodic cohomology if and only if its Sylow p-subgroups are cyclic, for all odd primes p, while its Sylow 2-subgroups are either cyclic or generalized quaternion (see Adem–Milgram, *Cohomology of Finite Groups*, p. 148). For example, $G = \text{SL}(2, 5)$ has periodic cohomology: it is a group of order $120 = 8 \cdot 3 \cdot 5$, so its Sylow 3-subgroups and its Sylow 5-subgroups are cyclic, having prime order, while its Sylow 2-subgroups are isomorphic to the quaternions. ◂

We can interpret $H^1(G, A)$ and $H^2(G, A)$, where G is any not necessarily cyclic group, in terms of derivations and extensions if we can show that the formulas in Section 10.3 do, in fact, arise from a projective resolution of $\mathbb{Z}$. Alas, we need a technical interlude.

Definition. If G is a group, define $B_0(G)$ to be the free G-module on the single generator $[\ ]$ (hence, $B_0(G) \cong \mathbb{Z}G$) and, for $n \geq 1$, define $B_n(G)$ to be the free G-module with basis all symbols $[x_1 \mid x_2 \mid \cdots \mid x_n]$, where $x_i \in G$. Define $\varepsilon \colon B_0(G) \to \mathbb{Z}$ by $\varepsilon([\ ]) = 1$ and, for $n \geq 1$, define $d_n \colon B_n(G) \to B_{n-1}(G)$ by

$$\begin{aligned} d_n \colon [x_1 \mid \cdots \mid x_n] \mapsto x_1[x_2 \mid \cdots \mid x_n] \\ + \sum_{i=1}^{n-1}(-1)^i [x_1 \mid \cdots \mid x_i x_{i+1} \mid \cdots \mid x_n] \\ + (-1)^n [x_1 \mid \cdots \mid x_{n-1}]. \end{aligned}$$

The ***bar resolution*** is the sequence

$$\mathbf{B}_\bullet(G) : \cdots \to B_2(G) \xrightarrow{d_2} B_1(G) \xrightarrow{d_1} B_0(G) \xrightarrow{\varepsilon} \mathbb{Z} \to 0.$$

Let us look at the low-dimensional part of the bar resolution.

$$\begin{aligned} &d_1 \colon [x] \mapsto x[\ ]; \\ &d_2 \colon [x \mid y] \mapsto x[y] - [xy] + [x]; \\ &d_3 \colon [x \mid y \mid z] \mapsto x[y \mid z] - [xy \mid z] + [x \mid yz] - [x \mid y] \end{aligned}$$

These are the formulas that arose in the earlier sections, but without the added conditions $[x \mid 1] = 0 = [1 \mid y]$ and $[1] = 0$. In fact, there are two bar resolutions; the bar resolution just defined, and another we shall soon see, called the *normalized bar resolution*.

The bar resolution is a free resolution of $\mathbb{Z}$, although it is not a routine calculation to see this; we prove that it is a resolution by comparing $\mathbf{B}_\bullet(G)$ to a resolution familiar to algebraic topologists.

Definition. If G is a group, let $P_n(G)$ be the free abelian group with basis all $(n+1)$-tuples of elements of G; make $P_n(G)$ into a G-module by defining

$$x(x_0, x_1, \ldots, x_n) = (xx_0, xx_1, \ldots, xx_n).$$

Define $\partial_n \colon P_n(G) \to P_{n-1}(G)$, whenever $n \geq 1$, by

$$\partial_n \colon (x_0, x_1, \ldots, x_n) \mapsto \sum_{i=0}^{n}(-1)^i (x_0, \ldots, \widehat{x_i}, \ldots, x_n),$$

where $\widehat{x_i}$ means that x_i has been deleted. $\mathbf{P}_\bullet(G)$ is called the ***homogenous resolution*** of $\mathbb{Z}$.

Note that $P_0(G)$ is the free abelian group with basis all (y), for $y \in G$, made into a G-module by $x(y) = (xy)$. In other words, $P_0(G) = \mathbb{Z}G$.

The proof that $\mathbf{P}_\bullet(G)$ is a projective resolution of $\mathbb{Z}$ will be broken into two parts.

Lemma 10.114. *The sequence*

$$\mathbf{P}_\bullet(G): \cdots \to P_2(G) \xrightarrow{\partial_2} P_1(G) \xrightarrow{\partial_1} P_0(G) \xrightarrow{\varepsilon} \mathbb{Z} \to 0,$$

where ε is the augmentation, is a complex.

Proof. It suffices to prove that $\partial_{n-1}\partial_n(x_0, x_1, \ldots, x_n) = 0$. Now

$$\begin{aligned}\partial_{n-1}\partial_n(x_0, x_1, \ldots, x_n) &= \sum_{i=0}^{n}(-1)^i \partial_{n-1}(x_0, \ldots, \widehat{x_i}, \ldots, x_n)\\ &= \sum_{i=0}^{n}(-1)^i \Big(\sum_{j<i}(-1)^j (x_0, \ldots, \widehat{x_j}, \ldots, \widehat{x_i}, \ldots, x_n)\Big)\\ &\quad + \sum_{i=0}^{n}(-1)^i \Big(\sum_{j>i}(-1)^{j-1} (x_0, \ldots, \widehat{x_i}, \ldots, \widehat{x_j}, \ldots, x_n)\Big).\end{aligned}$$

In the last equation, the first summation has inner sign $(-1)^j$, because $j < i$ and so x_j is still in the jth position after the deletion of x_i from the original n-tuple. In the second summation, however, the inner sign is $(-1)^{j-1}$, because $i < j$ and so x_j is in position $j-1$ after deletion of the earlier x_i. Thus, $\partial_{n-1}\partial_n(x_0, x_1, \ldots, x_n)$ is a sum of $(n-2)$-tuples $(x_0, \ldots, \widehat{x_i}, \ldots, \widehat{x_j}, \ldots, x_n)$ with $i < j$, each of which occurs twice: once upon deleting x_i by ∂_n and then deleting x_j by ∂_{n-1}; a second time upon deleting x_j by ∂_n and then deleting x_i by ∂_{n-1}. In the first case, the sign of the $(n-2)$-tuple is $(-1)^{i+j-1}$; in the second case, its sign is $(-1)^{i+j}$. Therefore, the $(n-2)$-tuples cancel in pairs, and $\partial_{n-1}\partial_n = 0$. •

Proposition 10.115. *The complex*

$$\mathbf{P}_\bullet(G): \cdots \to P_2(G) \xrightarrow{\partial_2} P_1(G) \xrightarrow{\partial_1} P_0(G) \xrightarrow{\partial_0} \mathbb{Z} \to 0,$$

where $\partial_0 = \varepsilon$ is the augmentation, is a G-free resolution of $\mathbb{Z}$.

Proof. We let the reader prove that $P_n(G)$ is a free G-module with basis all symbols of the form $(1, x_1, \ldots, x_n)$.

To prove exactness of $\mathbf{P}_\bullet(G)$, it suffices, by Proposition 10.40, to construct a contracting homotopy; that is, maps

$$\cdots \leftarrow P_2(G) \xleftarrow{s_1} P_1(G) \xleftarrow{s_0} P_0(G) \xleftarrow{s_{-1}} \mathbb{Z}$$

with $\varepsilon s_{-1} = 1_{\mathbb{Z}}$ and

$$\partial_{n+1}s_n + s_{n-1}\partial_n = 1_{P_n(G)}, \quad \text{for all } n \geq 0.$$

Define $s_{-1}\colon \mathbb{Z} \to P_0(G)$ by $m \mapsto m(1)$, where the 1 in the parentheses is the identity element of the group G, and, for $n \geq 0$, define $s_n\colon P_n(G) \to P_{n+1}(G)$ by

$$s_n\colon (x_0, x_1, \ldots, x_n) \mapsto (1, x_0, x_1, \ldots, x_n).$$

These maps s_n are only $\mathbb{Z}$-maps, but Exercise 10.32 on page 829 says that this suffices to prove exactness. Here are the computations.

$$\varepsilon s_{-1}(1) = \varepsilon\big((1)\big) = 1$$

If $n \geq 0$, then

$$\begin{aligned}\partial_{n+1} s_n(x_0, \ldots, x_n) &= \partial_{n+1}(1, x_0, \ldots, x_n)\\ &= (x_0, \ldots, x_n) + \sum_{i=0}^{n} (-1)^{i+1}(1, x_0, \ldots, \widehat{x_i}, \ldots, x_n)\end{aligned}$$

[the range of summation has been rewritten because x_i sits in the $(i+1)$st position in $(1, x_0, \ldots, x_n)$]. On the other hand,

$$\begin{aligned}s_{n-1}\partial_n(x_0, \ldots, x_n) &= s_{n-1} \sum_{j=0}^{n} (-1)^j (x_0, \ldots, \widehat{x_j}, \ldots, x_n)\\ &= \sum_{j=0}^{n} (-1)^j (1, x_0, \ldots, \widehat{x_j}, \ldots, x_n).\end{aligned}$$

It follows that $\big(\partial_{n+1} s_n + s_{n-1}\partial_n\big)(x_0, \ldots, x_n) = (x_0, \ldots, x_n)$. •

Proposition 10.116. *The bar resolution* $\mathbf{B}_\bullet(\mathbf{G})$ *is a* G*-free resolution of* $\mathbb{Z}$.

Proof. For each $n \geq 0$, define $\tau_n\colon P_n(G) \to B_n(G)$ by

$$\tau_n\colon (x_0, \ldots, x_n) \mapsto x_0[x_0^{-1}x_1 \mid x_1^{-1}x_2 \mid \cdots \mid x_{n-1}^{-1}x_n],$$

and define $\sigma_n\colon B_n(G) \to P_n(G)$ by

$$\sigma_n\colon [x_1 \mid \cdots \mid x_n] \mapsto (1, x_1, x_1x_2, x_1x_2x_3, \ldots, x_1x_2\cdots x_n).$$

It is routine to check that τ_n and σ_n are inverse, and so each τ_n is an isomorphism.

The reader can also check that $\tau\colon \mathbf{P}_\bullet(\mathbf{G}) \to \mathbf{B}_\bullet(\mathbf{G})$ is a chain map; that is, the following diagram commutes:

$$\begin{array}{ccc} P_n(G) & \xrightarrow{\tau_n} & B_n(G) \\ {\scriptstyle\partial_n}\downarrow & & \downarrow{\scriptstyle d_n} \\ P_{n-1}(G) & \xrightarrow{\tau_{n-1}} & B_{n-1}(G) \end{array}$$

Finally, Exercise 10.22 on page 827 shows that both complexes have the same homology groups. By Proposition 10.115, the complex $\mathbf{P}_\bullet(G)$ is an exact sequence, so that all its homology groups are $\{0\}$. It follows that all the homology groups of $\mathbf{B}_\bullet(G)$ are $\{0\}$, and so it, too, is an exact sequence. •

Definition. Define

$$[x_1 \mid \cdots \mid x_n]^* = \begin{cases} [x_1 \mid \cdots \mid x_n] & \text{if all } x_i \neq 1; \\ 0 & \text{if some } x_i = 1. \end{cases}$$

The ***normalized bar resolution***, $\mathbf{B}^*_\bullet(\mathbf{G})$, is the sequence

$$\mathbf{B}^*_\bullet(\mathbf{G}) : \cdots \to B^*_2(G) \xrightarrow{d_2} B^*_1(G) \xrightarrow{d_1} B^*_0(G) \xrightarrow{\varepsilon} \mathbb{Z} \to 0,$$

where $B^*_n(G)$ is the free G-module with basis all nonzero $[x_1 \mid \cdots \mid x_n]^*$, and the maps d_n have the same formula as the maps d_n in the bar resolution (except that the symbols $[x_1 \mid \cdots \mid x_n]$ now occur as $[x_1 \mid \cdots \mid x_n]^*$).

Since we are making some of the basis elements 0, it is not obvious that the normalized bar resolution $\mathbf{B}^*_\bullet(\mathbf{G})$ is a complex, let alone a resolution of $\mathbb{Z}$.

Theorem 10.117. *The normalized bar resolution $\mathbf{B}^*_\bullet(\mathbf{G})$ is a G-free resolution of $\mathbb{Z}$.*

Proof. We begin by constructing a contracting homotopy

$$\cdots \leftarrow B^*_2(G) \xleftarrow{t_1} B^*_1(G) \xleftarrow{t_0} B^*_0(G) \xleftarrow{t_{-1}} \mathbb{Z},$$

where each t_n is a $\mathbb{Z}$-map. Define $t_{-1} \colon \mathbb{Z} \to B^*_0(G)$ by $t_{-1} \colon m \mapsto m[\ ]$. Note that $B^*_n(G)$ is a free G-module with basis all nonzero $[x_1 \mid \cdots \mid x_n]^*$; hence, it is a direct sum of copies of $\mathbb{Z}G$. Since $\mathbb{Z}G$ is a free abelian group, $B^*_n(G)$ is also a free abelian group; the reader may check that a basis of $B^*_n(G)$, as a free abelian group, consists of all nonzero $x[x_1 \mid \cdots \mid x_n]^*$. To define t_n for $n \geq 0$, we take advantage of the fact that t_n need only be a $\mathbb{Z}$-map, by giving its values on these $\mathbb{Z}$-basis elements (and freeness allows us to choose these values without restrictions). Thus, for $n \geq 0$, define $t_n \colon B^*_n(G) \to B^*_{n+1}(G)$ by

$$t_n \colon x[x_1 \mid \cdots \mid x_n]^* \mapsto [x \mid x_1 \mid \cdots \mid x_n]^*.$$

That we have constructed a contracting homotopy is routine; the reader may check that $\varepsilon t_{-1} = 1_{\mathbb{Z}}$ and, for $n \geq 0$, that

$$d_{n+1} t_n + t_{n-1} d_n = 1_{B^*_n(G)}.$$

The proof will be complete once we show that $\mathbf{B}^*_\bullet(\mathbf{G})$ is a complex. Since $B^*_{n+1}(G)$ is generated, as a G-module, by $\operatorname{im} t_n$, it suffices to show that $d_n d_{n+1} = 0$ on this subgroup. We now prove, by induction on $n \geq -1$, that $d_n d_{n+1} t_n = 0$. The base step is true, for $\varepsilon = t_{-1}$ and $0 = \varepsilon d_1 = t_{-1} d_1$. For the inductive step, we use the identities in the definition of contracting homotopy and the inductive hypothesis $d_{n-1} d_n = 0$:

$$\begin{aligned} d_n d_{n+1} t_n &= d_n(1 - t_{n-1} d_n) \\ &= d_n - d_n t_{n-1} d_n \\ &= d_n - (1 - t_{n-2} d_{n-1}) d_n \\ &= d_n - d_n - t_{n-2} d_{n-1} d_n \\ &= 0. \quad \bullet \end{aligned}$$

We can now interpret $H^1(G, A)$ and $H^2(G, A)$.

Corollary 10.118. *For every group G and every G-module A, the groups $H^1(G, A)$ and $H^2(G, A)$ constructed in Section* 10.3 *coincide with the cohomology groups.*

Proof. We have proved that factor sets, coboundaries, derivations, and principal derivations do, in fact, arise from a projective resolution of $\mathbb{Z}$. •

Proposition 10.119. *If G is a finite group of order m, then $mH^n(G, A) = \{0\}$ for all $n \geq 1$ and all G-modules A.*

Proof. Sum the cocycle formula, as in the proof of Theorem 10.21. •

Corollary 10.120. *If G is a finite group and A is a finitely generated G-module, then $H^n(G, A)$ are finite for all $n \geq 0$.*

Proof. $H^n(G, A)$ is a finitely generated abelian group (because A is finitely generated) of finite exponent. •

Both Proposition 10.119 and its corollary are true for homology groups as well.

There are several aspects of the cohomology of groups that we have not mentioned. Aside from being a useful tool within group theory itself, these groups also form a link with algebraic topology. For every group G, there exists a topological space $K(G, 1)$, called an Eilenberg–Mac Lane space, whose fundamental group is G and whose cohomology groups coincide with the algebraically defined cohomology groups.[16] There is, in fact, a deep connection between group theory and algebraic topology, of which this is a first sign.

An important property of the cohomology of groups is the relation between the cohomology of a group and the cohomology of its subgroups and its quotient groups. If $\varphi\colon S \to G$ is a homomorphism, every G-module A becomes an S-module if we define $sa = \varphi(s)a$ for all $s \in S$ and $a \in A$. What is the connection between $H^n(S, A)$ and $H^n(G, A)$? What is the connection between $H_n(S, A)$ and $H_n(G, A)$? [There is also a connection between homology groups and cohomology groups: $H^n(G, A)^* \cong H_n(G, A^*)$, where $A^* = \text{Hom}_{\mathbb{Z}}(A, \mathbb{Q}/\mathbb{Z})$.]

There are three standard maps, which we will define in terms of the bar resolution. The first is ***restriction***. If S is a subgroup of a group G, then every function $f\colon B_n(G) \to A$ is defined on all $[x_1 \mid \cdots \mid x_n]$ with $x_i \in G$; of course, f is defined on all n-tuples of the form $[s_1 \mid \cdots \mid s_n]$ with $s_i \in S \subseteq G$, so that its restriction, which we denote by $f|S$, maps $B_n(S) \to A$. If f is an n-cocycle, let us write $\text{cls}\, f$ to denote its cohomology class: $\text{cls}\, f = f + \text{im}\, d_{n+1}$. Then

$$\text{Res}\colon H^n(G, A) \to H^n(S, A)$$

is defined by $\text{Res}(\text{cls}\, f) = \text{cls}(f|S)$. One result is that if G is a finite group, S_p is a Sylow p-subgroup, and $n \geq 1$, then $\text{Res}\colon H^n(G, A) \to H^n(S_p, A)$ is injective on the

[16]Because of this topological connection, many authors use the notation $H^n(\pi, \mathbb{Z})$ to denote cohomology groups, for π_1 is the standard notation for the fundamental group.

p-primary component of $H^n(G, A)$; thus, the cohomology of G is strongly influenced by the cohomology of its Sylow subgroups.

If $S \leq G$, there is a map

$$\text{Cor}\colon H^n(S, A) \to H^n(G, A)$$

in the reverse direction, called ***corestriction***, which is defined when S has finite index in G. We first define Cor in dimension 0; that is, $\text{Cor}^0\colon A^S \to A^G$, by $a \mapsto \sum_{t \in T} ta$, where T is a left transversal of S in G (of course, we must check that Cor^0 is a homomorphism that is independent of the choice of transversal). There is a standard way of extending a map in dimension 0 to maps in higher dimensions (essentially by dimension shifting), and if $[G : S] = m$, then

$$\text{Cor}^n \circ \text{Res}^n\colon H^n(G, A) \to H^n(G, A) = m;$$

that is, the composite is multiplication by m. Similarly, in homology, the map

$$\text{Cor}_0\colon A/\mathcal{G}A \to A/\mathcal{S}A,$$

defined by $a + \mathcal{G}A \mapsto \sum_{t \in T} t^{-1}a + \mathcal{S}A$, extends to maps in higher dimensions. When $n = 1$ and $A = \mathbb{Z}$, we have $\text{Cor}_1\colon H_1(G, \mathbb{Z}) \to H_1(S, \mathbb{Z})$; that is, $\text{Cor}_1\colon G/G' \to S/S'$. There is such a homomorphism well known to group theorists, called the ***transfer*** $V_{G \to S}$, and it turns out that $\text{Cor}_1 = V_{G \to S}$.

The third standard map is called ***inflation***. Suppose that N is a normal subgroup of a group G. If A is a G-module, then A^N is a G/N-module if we define $(gN)a = ga$ for $a \in A^N$ [if $gN = hN$, then $h = gx$ for some $x \in N$, and so $ha = (gx)a = g(xa) = ga$, because $xa = a$]. Define

$$\text{Inf}\colon H^n(G/N, A^N) \to H^n(G, A)$$

by $\text{cls}\, f \mapsto \text{cls}(f^\#)$, where

$$f^\#\colon [g_1 \mid \cdots \mid g_n] \mapsto f[g_1 N \mid \cdots \mid g_n N].$$

A useful result here is the ***five term exact sequence***: If $N \lhd G$ and A is a G-module, there is an exact sequence of abelian groups:

$$\begin{aligned} 0 \to H^1(G/N, A^N) \xrightarrow{\text{Inf}} H^1(G, A) \xrightarrow{\text{Res}} H^1(N, A)^{G/N} \\ \longrightarrow H^2(G/N, A^N) \longrightarrow H^2(G, A) \end{aligned}$$

(there is a version of this sequence in homology as well). For an excellent discussion of these ideas, we refer the reader to Serre, *Corps Locaux*, pp. 135–138.

We can force cohomology groups to be a graded ring by defining ***cup product*** on $H^\bullet(G, R) = \sum_{n \geq 0} H^n(G, R)$, where R is any commutative ring (see Evens, *The Cohomology of Groups*), and this added structure has important applications.

We now consider the cohomology of free groups.

Lemma 10.121. *If G is a free group with basis X, then its augmentation ideal $\mathcal{G}$ is a free G-module with basis*

$$X - 1 = \{x - 1 : x \in X\}.$$

Proof. We show first that $\mathcal{G}$ is generated by all $X - 1$. The identities

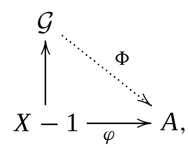
$$xy - 1 = (x - 1) + x(y - 1)$$

and

$$x^{-1} - 1 = -x^{-1}(x - 1)$$

show that if w is any word in X, then $w - 1$ can be written as a G-linear combination of $X - 1$.

To show that $\mathcal{G}$ is a free G-module with basis $X - 1$, it now suffices to show that the following diagram can be completed:



where A is any G-module and φ is any function (uniqueness of such a map Φ follows from $X - 1$ generating $\mathcal{G}$). Thus, we are seeking $\Phi \in \operatorname{Hom}_G(\mathcal{G}, A)$. By Exercise 10.59 on page 886, we have $\operatorname{Hom}_G(\mathcal{G}, A) \cong \operatorname{Der}(G, A)$ via $f : x \mapsto f(x - 1)$, where $f \in \mathcal{G} \to A$, and so we seek a derivation.

Consider the (necessarily split) extension $0 \to A \to E \to G \to 1$, so that E consists of all ordered pairs $(g, a) \in G \times A$. The given function $\varphi \colon X - 1 \to A$ defines a lifting ℓ of the generating set X of G, namely,

$$\ell(x) = (\varphi(x - 1), x).$$

Since G is free with basis X, the function $\ell \colon X \to E$ extends to a homomorphism $L : G \to E$. We claim, for every $g \in G$, that $L(g) = (d(g), g)$, where $d \colon G \to A$. Each $g \in G$ has a unique expression as a reduced word $g = x_1^{e_1} \cdots x_n^{e_n}$, where $x_i \in X$ and $e_i = \pm 1$. We prove the claim by induction on $n \geq 1$. The base step is clear, while

$$\begin{aligned} L(g) &= L(x_1^{e_1} \cdots x_n^{e_n}) \\ &= L(x_1^{e_1}) \cdots L(x_n^{e_n}) \\ &= (\varphi(x_1 - 1), x_1)^{e_1} \cdots (\varphi(x_n - 1), x_n)^{e_n} \\ &= (d(g), g), \end{aligned}$$

and so the first coordinate $d(g)$ lies in A.

Exercise 10.59 on page 886 now says that there is a homomorphism $\Phi \colon \mathcal{G} \to A$ defined by $\Phi(g - 1) = d(g)$ for all $g \in G$. In particular, $\Phi(x - 1) = d(x) = \varphi(x - 1)$, so that Φ does extend φ. •

Theorem 10.122. *If G is a free group, then $H^n(G, A) = \{0\}$ for all $n > 1$ and all G-modules A.*

Proof. The sequence $0 \to \mathcal{G} \to \mathbb{Z}G \to \mathbb{Z} \to 0$ is a free resolution of $\mathbb{Z}$ because $\mathcal{G}$ is now a free G-module. Thus, the only nonzero terms in the deleted resolution occur in positions 0 and 1, and so all cohomology groups vanish for $n > 1$. •

We are now going to state an interesting result (the Stallings–Swan theorem), which was discovered using homomological methods but which does not mention homology in its statement.

If G is a group and $S \le G$ is a subgroup, then every G-module A can be viewed as an S-module, for $\mathbb{Z}S$ is a subring of $\mathbb{Z}G$.

Definition. A group G has ***cohomological dimension*** $\le n$, in symbols, $\mathrm{cd}(G) \le n$, if

$$H^{n+1}(S, A) = \{0\}$$

for all G-modules A and every subgroup S of G. We write $\mathrm{cd}(G) = \infty$ if no such integer n exists.

We say that $\mathrm{cd}(G) = n$ if $\mathrm{cd}(G) \le n$ but it is not true that $\mathrm{cd}(G) \le n - 1$.

Example 10.123.

(i) If $G = \{1\}$, then $\mathrm{cd}(G) = 0$; this follows from Theorem 10.112 because G is a cyclic group of order 1.

(ii) If G is a finite cyclic group of order $k > 1$, then $\mathrm{cd}(G) = \infty$, as we see from Corollary 10.113 with $A = \mathbb{Z}$.

(iii) If $G \ne \{1\}$ is a free group, then Theorem 10.122 shows that $\mathrm{cd}(G) = 1$, for every subgroup of a free group is free.

(iv) If $\mathrm{cd}(G) < \infty$, then G must be torsion-free; otherwise, G has a subgroup S that is cyclic of finite order $k > 1$, and $H^n(S, \mathbb{Z}) \ne 0$ for all even n.

(v) It is known that if G is a free abelian group of finite rank n, then $\mathrm{cd}(G) = n$. ◂

Proposition 10.124 (Shapiro's Lemma). *Let G be a group and let $S \le G$ be a subgroup. If A is a $\mathbb{Z}S$-module, then for all $n \ge 0$,*

$$H^n(S, A) \cong H^n(G, \mathrm{Hom}_{\mathbb{Z}S}(\mathbb{Z}G, A)).$$

Proof. Let $\cdots \to P_1 \to P_0 \to \mathbb{Z} \to 0$ be a $\mathbb{Z}G$-free resolution. If we denote $\mathrm{Hom}_{\mathbb{Z}S}(\mathbb{Z}G, A))$ by A^*, then

$$H^n(G, A^*) = H^n(\mathrm{Hom}_{\mathbb{Z}G}(\mathbf{P}_\bullet, A^*)).$$

By the adjoint isomorphism,

$$\begin{aligned}\mathrm{Hom}_{\mathbb{Z}G}(P_i, A^*) &= \mathrm{Hom}_{\mathbb{Z}G}(P_i, \mathrm{Hom}_{\mathbb{Z}S}(\mathbb{Z}G, A)) \\ &\cong \mathrm{Hom}_{\mathbb{Z}S}(P_i \otimes_{\mathbb{Z}G} \mathbb{Z}G, A) \\ &\cong \mathrm{Hom}_{\mathbb{Z}S}(P_i, A).\end{aligned}$$

But (the proof of) Lemma 8.141(i) shows that $\mathbb{Z}G$ is a free $\mathbb{Z}S$-module, and so the free $\mathbb{Z}G$-modules P_i are also free $\mathbb{Z}S$-modules. It follows that we may regard $\mathbf{P}_\bullet$ as a $\mathbb{Z}S$-free resolution of $\mathbb{Z}$, and there is an isomorphism of complexes:

$$\mathrm{Hom}_{\mathbb{Z}S}(\mathbf{P}_\bullet, A) \cong \mathrm{Hom}_{\mathbb{Z}G}(\mathbf{P}_\bullet, A^*).$$

Hence, their homology groups are isomorphic; that is, $H^n(S, A) \cong H^n(G, A^*)$. •

Corollary 10.125. *If G is a group and $S \le G$ is a subgroup, then*

$$\mathrm{cd}(S) \le \mathrm{cd}(G).$$

Proof. We may assume that $\mathrm{cd}(G) = n < \infty$. If $m > n$ and there is a $\mathbb{Z}S$-module A with $H^m(S, A) \neq \{0\}$, then Shapiro's lemma gives $H^m(G, \mathrm{Hom}_{\mathbb{Z}S}(\mathbb{Z}G, A)) \cong H^m(S, A) \neq \{0\}$, and this contradicts $\mathrm{cd}(G) = n$. •

Corollary 10.126. *A group G of finite cohomological dimension has no elements* (*other than* 1) *of finite order.*

Proof. This follows at once from Example 10.123(ii) and the preceding corollary. •

Are there groups G with $\mathrm{cd}(G) = 1$ that are not free? In 1970, J. Stallings proved the following nice theorem ($\mathbb{F}_2G$ denotes the group algebra over $\mathbb{F}_2$).

Theorem. *If G is a finitely presented group for which $H^1(G, \mathbb{F}_2G)$ has more than* 2 *elements, then G is a free product, $G = H * K$, where $H \neq \{1\}$ and $K \neq \{1\}$* (*free product is the coproduct in* **Groups**).

As a consequence, he proves the following results.

Corollary. *If G is a finitely generated group with* $\mathrm{cd}(G) = 1$, *then G is free.*

Corollary. *If G is a torsion-free finitely generated group having a free subgroup of finite index, then G is free.*

R. G. Swan showed that both corollaries remain true if we remove the hypothesis that G be finitely generated.

Theorem (Stallings–Swan). *A torsion-free group having a free subgroup of finite index must be free.*

Corollary 10.127 (Nielsen–Schreier). *Every subgroup S of a free group F is itself free.*

Proof. By Corollary 10.125, we have $\mathrm{cd}(S) \leq 1$, and so the result follows from the theorem of Stallings and Swan. •

We refer the reader to D. E. Cohen, "Groups of Cohomological Dimension 1," *Lecture Notes in Mathematics*, Vol. 245, Springer–Verlag, New York, 1972, for proofs of these theorems.

EXERCISES

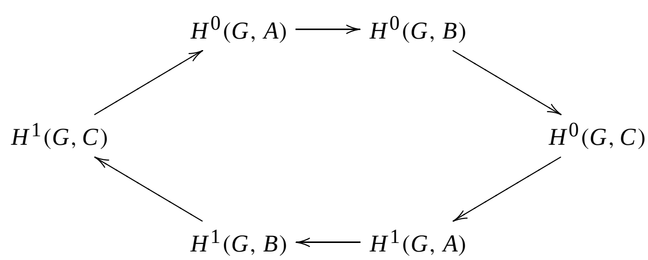
10.58 (i) Prove that the isomorphisms in Proposition 10.104 constitute a natural equivalence $\mathbb{Z}\otimes_G$ to $A \mapsto A/\mathcal{G}A$.

(ii) Prove that the isomorphisms in Proposition 10.111 constitute a natural equivalence $\mathrm{Hom}_G(\mathbb{Z},\ )$ to $A \mapsto A^G$.

10.59 For a fixed group G, prove that the functors $\mathrm{Hom}_G(\mathcal{G},\ )$ and $\mathrm{Der}(G,\ )$ are naturally equivalent.

Hint. If $f : \mathcal{G} \to A$ is a homomorphism, then $d_f : x \mapsto f(x-1)$ is a derivation.

10.60 (i) If G is a finite cyclic group and $0 \to A \to B \to C \to 0$ is an exact sequence of G-modules, prove that there is an ***exact hexagon***; that is, kernel = image at each vertex of the diagram



We remark that this exercise is a key lemma in class field theory.

(ii) If G is a finite cyclic group and A is a G-module, define the ***Herbrand quotient*** by

$$h(A) = |H^0(G, A)|/|H^1(G, A)|$$

[$h(A)$ is defined only when both $H^0(G, A)$ and $H^1(G, A)$ are finite].

Let $0 \to A \to B \to C \to 0$ be an exact sequence of G-modules. Prove that if the Herbrand quotient is defined for two of the modules A, B, C, then it is defined for the third one, and

$$h(B) = h(A)h(C).$$

10.61 If G is a group, prove that

$$P_n(G) \cong \bigotimes_1^{n+1} \mathbb{Z}G,$$

where $P_n(G)$ is the nth term in the homogeneous resolution $\mathbf{P}_\bullet(G)$ and

$$\bigotimes_1^n \mathbb{Z}G = \mathbb{Z}G \otimes_\mathbb{Z} \mathbb{Z}G \otimes_\mathbb{Z} \cdots \otimes_\mathbb{Z} \mathbb{Z}G,$$

the tensor product over $\mathbb{Z}$ of $\mathbb{Z}G$ with itself n times.

10.62 If G is a finite cyclic group, prove, for all G-modules A and for all $n \geq 1$, that $H^n(G, A) \cong H_{n+1}(G, A)$.

10.63 Let G be a group.

(i) Show, for any abelian group A, that $A^* = \operatorname{Hom}_\mathbb{Z}(\mathbb{Z}G, A)$ is a left $\mathbb{Z}G$-module. We call A^* a ***coinduced module***.

Hint. If $\varphi\colon \mathbb{Z}G \to A$ and $g \in G$, define $g\varphi$ by $x \mapsto g\varphi(g^{-1}x)$.

(ii) For any left $\mathbb{Z}G$-module B, prove that $\operatorname{Hom}_{\mathbb{Z}G}(B, A^*) \cong \operatorname{Hom}_\mathbb{Z}(B, A)$.

Hint. Use the adjoint isomorphism, Theorem 8.99.

(iii) If A^* is a coinduced module, prove that $H^n(G, A^*) = \{0\}$ for all $n \geq 1$.

10.64 If G is a group and A is an abelian group, call the $\mathbb{Z}G$-module $A_*\mathbb{Z}G \otimes_\mathbb{Z} A$ an ***induced module***. Prove that $H_n(G, A_*) = \{0\}$ for all $n \geq 1$.

10.8 Crossed Products

This section is essentially descriptive, showing how cohomology groups are used to study division rings. Let us begin with a return to Galois theory.

Theorem 10.128. *Let E/k be a Galois extension with Galois group $G = \operatorname{Gal}(E/k)$. The multiplicative group $E^\times$ is a kG-module, and*

$$H^1(G, E^\times) = \{0\}.$$

Proof. If $c\colon G \to E^\times$ is a 1-cocycle, denote $c(\sigma)$ by c_σ. In multiplicative notation, the cocycle condition is the identity $\sigma(c_\tau)c_{\sigma\tau}^{-1}c_\sigma = 1$ for all $\sigma, \tau \in G$; that is,

$$\sigma(c_\tau) = c_{\sigma\tau}c_\sigma^{-1}. \tag{1}$$

For $e \in E^\times$, consider

$$b = \sum_{\tau \in G} c_\tau \tau(e).$$

By independence of characters, Proposition 4.30, there is some $e \in E^\times$ with $b \neq 0$. For

such an element e, we have, using Eq. (1),

$$\begin{aligned}\sigma(b) &= \sum_{\tau \in G} \sigma(c_\tau)\sigma\tau(e) \\ &= \sum_{\tau \in G} c_{\sigma\tau} c_\sigma^{-1} \sigma\tau(e) \\ &= c_\sigma^{-1} \sum_{\tau \in G} c_{\sigma\tau}\sigma\tau(e) \\ &= c_\sigma^{-1} \sum_{\omega \in G} c_\omega \omega(e) \\ &= c_\sigma^{-1} b.\end{aligned}$$

Hence, $c_\sigma = b\sigma(b)^{-1}$, and c is a coboundary. Therefore, $H^1(G, E^\times) = \{0\}$. •

Theorem 10.128 implies Theorem 4.50, which describes the elements of norm 1 in a cyclic extension.

Corollary 10.129 (Hilbert's Theorem 90). *Let E/k be a Galois extension whose Galois group $G = \text{Gal}(E/k)$ is cyclic, say, with generator σ. If $u \in E^\times$, then $Nu = 1$ if and only if there is $v \in E^\times$ with*

$$u = \sigma(v)v^{-1}.$$

Proof. By Theorem 10.112, we have $H^1(G, E^\times) = \ker N / \operatorname{im} D$, where N is the norm (remember that $E^\times$ is a multiplicative group) and $De = \sigma(e)e^{-1}$. Theorem 10.128 gives $H^1(G, E^\times) = \{0\}$, so that $\ker N = \operatorname{im} D$. Hence, if $u \in E^\times$, then $Nu = 1$ if and only if there is $v \in E^\times$ with $u = \sigma(v)v^{-1}$. •

Theorem 10.128 is one of the first results in what is called *Galois cohomology*. Another early result is that $H^n(G, E) = \{0\}$ for all $n \geq 1$, where E (in contrast to $E^\times$) is the additive group of the Galois extension (this result follows easily from the normal basis theorem). We are now going to see that $H^2(G, E^\times)$ is useful in studying division rings.

Only one example of a noncommutative division ring has been given in the text: the quaternions $\mathbb{H}$ (this is an $\mathbb{R}$-algebra) and its k-algebra analogs for every subfield $k \subseteq \mathbb{R}$ (actually, another example is given in Exercise 9.80 on page 740). W. R. Hamilton discovered the quaternions in 1843, and F. G. Frobenius, in 1880, proved that the only $\mathbb{R}$-division algebras are $\mathbb{R}$, $\mathbb{C}$, and $\mathbb{H}$ (see Theorem 9.124). No other examples of noncommutative division rings were known until *cyclic algebras* were found in the early 1900s, by J. M. Wedderburn and by L. E. Dickson. In 1932, A. A. Albert found an example of a *crossed product algebra* that is not a cyclic algebra, and in 1972, S. A. Amitsur found an example of a noncommutative division ring that is not a crossed product algebra.

Wedderburn proved that every finite division ring is a field (see Theorem 8.23). Are there any division rings of prime characteristic?

We begin with an elementary calculation. Suppose that V is a vector space over a field E having basis $\{u_\sigma : \sigma \in G\}$ for some set G, so that each $v \in V$ has a unique expression

as an E-linear combination $v = \sum_\sigma a_\sigma u_\sigma$ for $a_\sigma \in E$. For a function $\mu \colon V \times V \to V$, with $\mu(u_\sigma, u_\tau)$ denoted by $u_\sigma u_\tau$, define ***structure constants*** $g_\alpha^{\sigma,\tau} \in E$ by

$$u_\sigma u_\tau = \sum_{\alpha \in G} g_\alpha^{\sigma,\tau} u_\alpha.$$

To have the associative law, we must have $u_\sigma(u_\tau u_\omega) = (u_\sigma u_\tau)u_\omega$; expanding this equation gives, for all indices,

$$\sum_{\alpha,\beta} g_\alpha^{\sigma,\tau} g_\beta^{\alpha,\omega} = \sum_{\gamma,\delta} g_\gamma^{\tau,\omega} g_\delta^{\sigma,\gamma}.$$

Let us simplify these equations. Let G be a group and suppose that $g_\alpha^{\sigma,\tau} = 0$ unless $\alpha = \sigma\tau$; that is, $u_\sigma u_\tau = f(\sigma,\tau)u_{\sigma\tau}$, where $f(\sigma,\tau) = g_{\sigma\tau}^{\sigma,\tau}$. The function $f : G \times G \to E^\times$, given by $f(\sigma,\tau) = g_{\sigma\tau}^{\sigma,\tau}$, satisfies the following equation for all $\sigma, \tau, \omega \in G$:

$$f(\sigma,\tau) f(\sigma\tau,\omega) = f(\tau,\omega) f(\sigma,\tau\omega),$$

an equation reminiscent of the cocycle identity written in multiplicative notation. This is why factor sets enter into the next definition.

Let E/k be a Galois extension with $\mathrm{Gal}(E/k) = G$, and let $f : G \times G \to E^\times$ be a factor set: In multiplicative notation

$$f(\sigma, 1) = 1 = f(1, \tau) \quad \text{for all } \sigma, \tau \in G$$

and, if we denote the action of $\sigma \in G$ on $a \in E^\times$ by a^σ, then

$$f(\sigma,\tau) f(\sigma\tau,\omega) = f(\tau,\omega)^\sigma f(\sigma,\tau\omega).$$

Definition. Given a Galois extension E/k with Galois group $G = \mathrm{Gal}(E/k)$ and a factor set $f : G \times G \to E^\times$, define the ***crossed product algebra*** (E, G, f) to be the vector space over E having basis all symbols $\{u_\sigma : \sigma \in G\}$ and multiplication

$$(au_\sigma)(bu_\tau) = ab^\sigma f(\sigma,\tau)u_{\sigma\tau}$$

for all $a, b \in E$. If G is a cyclic group, then the crossed product algebra (E, G, f) is called a ***cyclic algebra***.

Since every element in (E, G, f) has a unique expression of the form $\sum a_\sigma u_\sigma$, the definition of multiplication extends by linearity to all of (E, G, f). We note two special cases:

$$u_\sigma b = b^\sigma u_\sigma;$$
$$u_\sigma u_\tau = f(\sigma,\tau)u_{\sigma\tau}.$$

Proposition 10.130. *If E/k is a Galois extension with Galois group $G = \mathrm{Gal}(E/k)$ and if $f : G \times G \to E^\times$ is a factor set, then (E, G, f) is a central simple k-algebra that is split by E.*

Proof. Denote (E, G, f) by A. First, we show that A is a k-algebra. To prove that A is associative, it suffices to prove that

$$au_\sigma(bu_\tau cu_\omega) = (au_\sigma bu_\tau)cu_\omega,$$

where $a, b, c \in E$. Using the definition of multiplication,

$$\begin{aligned} au_\sigma(bu_\tau cu_\omega) &= au_\sigma(bc^\tau f(\tau,\omega)u_{\tau\omega}) \\ &= a\big(bc^\tau f(\tau,\omega)\big)^\sigma f(\sigma,\tau\omega)u_{\sigma\tau\omega} \\ &= ab^\sigma c^{\sigma\tau} f(\tau,\omega)^\sigma f(\sigma,\tau\omega)u_{\sigma\tau\omega}. \end{aligned}$$

We also have

$$\begin{aligned} (au_\sigma bu_\tau)cu_\omega &= ab^\sigma f(\sigma,\tau)u_{\sigma\tau}cu_\omega \\ &= ab^\sigma f(\sigma,\tau)c^{\sigma\tau} f(\sigma\tau,\omega)u_{\sigma\tau\omega} \\ &= ab^\sigma c^{\sigma\tau} f(\sigma,\tau) f(\sigma\tau,\omega)u_{\sigma\tau\omega}. \end{aligned}$$

The cocycle identity shows that multiplication in A is associative.

That u_1 is the unit in A follows from our assuming that factor sets are normalized:

$$u_1 u_\tau = f(1,\tau)u_{1\tau} = u_\tau \quad \text{and} \quad u_\sigma u_1 = f(\sigma,1)u_{\sigma 1} = u_\sigma.$$

We have shown that A is a ring. We claim that $ku_1 = \{au_1 : a \in k\}$ is the center $Z(A)$. If $a \in E$, then $u_\sigma a u_1 = a^\sigma u_\sigma$. If $a \in k = E^G$, then $a^\sigma = a$ for all $\sigma \in G$, and so $k \subseteq Z(A)$. For the reverse inclusion, suppose that $z = \sum_\sigma a_\sigma u_\sigma \in Z(A)$. For any $b \in E$, we have $zbu_1 = bu_1 z$. But

$$zbu_1 = \sum a_\sigma u_\sigma bu_1 = \sum a_\sigma b^\sigma u_\sigma.$$

On the other hand,

$$bu_1 z = \sum ba_\sigma u_\sigma.$$

For every $\sigma \in G$, we have $a_\sigma b^\sigma = ba_\sigma$, so that if $a_\sigma \neq 0$, then $b^\sigma = b$. If $\sigma \neq 1$ and $H = \langle\sigma\rangle$, then $E^H \neq E^{\{1\}} = E$, by Theorem 4.33, and so there exists $b \in E$ with $b^\sigma \neq b$. We conclude that $z = a_1 u_1$. For every $\sigma \in G$, the equation $(a_1 u_1)u_\sigma = u_\sigma(a_1 u_1)$ gives $a_1^\sigma = a_1$, and so $a_1 \in E^G = k$. Therefore, $Z(A) = ku_1$.

We now show that A is simple. Let I be a nonzero two-sided ideal in A, and choose a nonzero $y \in I$ of shortest length; that is, $y = \sum c_\sigma u_\sigma$ has the smallest number of nonzero coefficients c_σ. Suppose that $y = c_\sigma u_\sigma + c_\tau u_\tau + \cdots$ has at least two nonzero coefficients c_σ and c_τ. Since $u_\tau u_{\tau^{-1}\sigma} = f(\tau, \tau^{-1}\sigma)u_\sigma$, it follows that $yu_{\tau^{-1}\sigma} \in I$, and that $y - c_\sigma c_\tau^{-1} f(\tau,\tau^{-1}\sigma)^{-1} yu_{\tau^{-1}\sigma}$ is an element of I whose length is shorter than that of y

(this element is nonzero because the coefficient of $u_{\sigma\tau^{-1}\sigma}$ is nonzero). We conclude that y has length 1; that is, $y = c_\sigma u_\sigma$. Hence, I contains $c_\sigma^{-1} f(\sigma, \sigma^{-1})^{-1}(c_\sigma u_\sigma)u_{\sigma^{-1}} = u_1$, the identity of A. Therefore, $I = A$, and A is simple.

Finally, Theorem 9.127 says that A is split by K, where K is any maximal subfield of A. The reader may show, using Lemma 9.117, that $Eu_1 \cong E$ is a maximal subfield. •

In light of Proposition 10.130, it is natural to expect a connection between relative Brauer groups and cohomology.

Theorem. *Let E/k be a Galois extension with $G = \mathrm{Gal}(E/k)$. There is an isomorphism $H^2(G, E^\times) \to \mathrm{Br}(E/k)$ with $\mathrm{cls}\, f \mapsto [(G, E, f)]$.*

Sketch of Proof. The usual proofs of this theorem are rather long. Each of the items: the isomorphism is a well-defined function; it is a homomorphism; it is injective; it is surjective, must be checked, and the proofs are computational. For example, the proof in Herstein, *Noncommutative Rings* covers pages 110 through 116. There is a less computational proof in Serre, *Corps Locaux*, pages 164 – 167, using the method of *descent*. •

What is the advantage of this isomorphism? In Corollary 9.132, we saw that

$$\mathrm{Br}(k) = \bigcup_{E/k \text{ finite}} \mathrm{Br}(E/k).$$

Corollary 10.131. *Let k be a field.*

(i) *The Brauer group* $\mathrm{Br}(k)$ *is a torsion group.*

(ii) *If A is a central simple k-algebra, then there is an integer n so that the tensor product of A with itself r times (where r is the order of $[A]$ in $\mathrm{Br}(k)$) is a matrix algebra*:

$$A \otimes_k A \otimes_k \cdots \otimes_k A \cong \mathrm{Mat}_n(k).$$

Sketch of Proof. (i) $\mathrm{Br}(k)$ is the union of the relative Brauer groups $\mathrm{Br}(E/k)$, where E/k is finite. It can be shown that $\mathrm{Br}(k)$ is the union of those $\mathrm{Br}(E/k)$ for which E/k is a Galois extension. We may now invoke Proposition 10.119, which says that $|G|\, H^2(G, E^\times) = \{0\}$.
(ii) Tensor product is the binary operation in the Brauer group. •

Recall Proposition 9.129: there exists a noncommutative division k-algebra over a field k if and only if $\mathrm{Br}(k) \neq \{0\}$.

Corollary 10.132. *Let k be a field. If there is a cyclic Galois extension E/k such that the norm $N\colon E^\times \to k^\times$ is not surjective, then there exists a noncommutative k-division algebra.*

Sketch of Proof. If G is a finite cyclic group, then Theorem 10.112 gives

$$H^2(G, E^\times) = (E^\times)^G/\operatorname{im} N = k^\times/\operatorname{im} N.$$

Therefore, $\mathrm{Br}(E/k) \neq \{0\}$ if N is not surjective, and this implies that $\mathrm{Br}(k) \neq \{0\}$. •

If k is a finite field and E/k is a finite extension, then it follows from Wedderburn's theorem on finite division rings (Theorem 8.23) that the norm $N\colon E^\times \to k^\times$ is surjective.

Corollary 10.133. *If p is a prime, then there exists a noncommutative division algebra of characteristic p.*

Proof. If k is a field of characteristic p, it suffices to find a cyclic extension E/k for which the norm $N\colon E^\times \to k^\times$ is not surjective; that is, we must find some $z \in k^\times$ which is not a norm.

If p is an odd prime, let $k = \mathbb{F}_p(x)$. Since p is odd, $t^2 - x$ is a separable irreducible polynomial, and so $E = k(\sqrt{x})$ is a Galois extension of degree 2. If $u \in E$, then there are polynomials $a, b, c \in \mathbb{F}_p[x]$ with $u = (a + b\sqrt{x})/c$. Moreover,

$$N(u) = (a^2 - b^2 x)/c^2.$$

We claim that $x^2 + x$ is not a norm. Otherwise,

$$a^2 - b^2 x = c^2(x^2 + x).$$

Since $c \neq 0$, the polynomial $c^2(x^2 + x) \neq 0$, and it has even degree. On the other hand, if $b \neq 0$, then $a^2 - b^2 x$ has odd degree, and this is a contradiction. If $b = 0$, then $u = a/c$; since $a^2 = c^2(x^2 + x)$, we have $c^2 \mid a^2$, hence $c \mid a$, and so $u \in \mathbb{F}_p[x]$ is a polynomial. But it is easy to see that $x^2 + x$ is not the square of a polynomial. We conclude that $N\colon E^\times \to k^\times$ is not surjective.

Here is an example in characteristic 2. Let $k = \mathbb{F}_2(x)$, and let $E = k(\alpha)$, where α is a root of $f(t) = t^2 + t + x + 1$ [$f(t)$ is irreducible and separable; its other root is $\alpha + 1$]. As before, each $u \in E$ can be written in the form $u = (a + b\alpha)/c$, where $a, b, c \in \mathbb{F}_2[x]$. Of course, we may assume that x is not a divisor of all three polynomials a, b and c. Moreover,

$$N(u) = \big((a + b\alpha)(a + b\alpha + b)\big)/c^2 = \big(a^2 + ab + b^2(x + 1)\big)/c^2.$$

We claim that x is not a norm. Otherwise,

$$a^2 + ab + b^2(x + 1) = c^2 x. \tag{2}$$

Now $a(0)$, the constant term of a, is either 0 or 1. Consider the four cases arising from the constant terms of a and b; that is, evaluate Eq. (2) at $x = 0$. We see that $a(0) = 0 = b(0)$; that is $x \mid a$ and $x \mid b$. Hence, $x^2 \mid a^2$ and $x^2 \mid b^2$, so that Eq. (2) has the form $x^2 d = c^2 x$, where $d \in \mathbb{F}_2[x]$. Dividing by x gives $xd = c^2$, which forces $c(0) = 0$; that is, $x \mid c$, and this is a contradiction. •

For further discussion of the Brauer group, see the article by Serre in Cassels–Fröhlich, *Algebraic Number Theory*, Jacobson, *Basic Algebra* II, pages 471–481, Reiner, *Maximal Orders*, Chapters 5, 7, and 8, and the article by V. P. Platonov and V. I. Yanchevskii, *Finite-Dimensional Division Algebras*, in Kostrikin–Shafarevich, *Encyclopaedia of Mathematical Sciences*, *Algebra* IX. In particular, a ***global field*** is a field which is either an arithmetic

number field [i.e., a finite extension of $\mathbb{Q}$] or a *function field* [a finite extension of $k(x)$, where k is a finite field]. To each global field, we assign a family of *local fields*. These fields are best defined in terms of discrete valuations. A ***discrete valuation*** on a field L is a function $v\colon L^\times \to \mathbb{N}$ such that, for all $a, b \in L$,

$$\begin{aligned} v(a) &= 0 \quad \text{if and only if } a = 0; \\ v(ab) &= v(a)v(b); \\ v(a+b) &= \max\{v(a), v(b)\}. \end{aligned}$$

Now $R = \{a \in L : v(a) \leq 1\}$ is a domain and $P = \{a \in L : v(a) < 1\}$ is a maximal ideal in R. We call R/P the ***residue field*** of L with respect to the discrete valuation v. A ***local field*** is a field which is complete with respect to the metric arising from a discrete valuation on it, and whose residue field is finite. It turns out that every local field is either a finite extension of $\mathbb{Q}_p$, the p-adic numbers (which is the fraction field of the p-adic integers $\mathbb{Z}_p$) or it is isomorphic to $\mathbb{F}_q[[x]]$, the ring of formal power series in one variable over a finite field $\mathbb{F}_q$. If k is a local field, then $\mathrm{Br}(k) \cong H^2(k_s, k^\times)$, where k_s/k is the maximal separable extension of k in the algebraic closure $\overline{k}$. If A is a central simple K-algebra, where K is a global field, and if K_v is a local field of K, then $K_v \otimes_K A$ is a central simple K_v-algebra. The ***Hasse–Brauer–Noether–Albert theorem*** states that if A is a central simple algebra over a global field K, then $A \cong K$ if and only if $K_v \otimes_K A \cong K_v$ for all associated local fields K_v. We merely mention that these results where used by C. Chevalley to develop *class field theory* [the branch of algebraic number theory involving Galois extensions (of possibly infinite degree) having abelian Galois groups]. See Neukirch–Schmidt–Wingberg, *Cohomology of Number Fields*.

For generalizations of the Brauer group [e.g., $\mathrm{Br}(k)$, where k is a commutative ring] and ties to Morita theory, see Orzech–Small, *The Brauer Group of Commutative Rings*. and Caenepeel, *Brauer Groups, Hopf Algebras, and Galois Theory*.

EXERCISES

10.65 Show that the structure constants in the crossed product (E, G, f) are

$$g_\alpha^{\sigma,\tau} = \begin{cases} f(\sigma,\tau) & \text{if } \alpha = \sigma\tau; \\ 0 & \text{otherwise.} \end{cases}$$

10.66 Prove that $\mathbb{H} \otimes_{\mathbb{R}} \mathbb{H} \cong \mathrm{Mat}_4(\mathbb{R})$.

10.9 INTRODUCTION TO SPECTRAL SEQUENCES

The last topic we mention is spectral sequences, whose major uses are in computing homology groups and in comparing homology groups of composites of functors. This brief section merely describes the setting for spectral sequences, in the hope that it will ease the

reader's first serious encounter with them. For a more complete account, we refer the reader to Mac Lane, *Homology*, Chapter XI, McCleary, *User's Guide to Spectral Sequences*, or Rotman, *An Introduction to Homological Algebra*, Chapter 11.

Call a series of submodules of a module K,

$$K = K_0 \supseteq K_1 \supseteq K_2 \supseteq \cdots \supseteq K_\ell = \{0\},$$

a ***filtration*** (instead of a normal series), and call the quotients K_i/K_{i+1} the ***factor modules*** of the filtration. We know that a module K may not be determined by the factor modules of a filtration; on the other hand, knowledge of the factor modules does give some information about K. For example, if all the factor modules are zero, then $K = \{0\}$; if all the factor modules are finite, then K is finite (and $|K|$ is the product of the orders of the factor modules); or, if all the factor modules are finitely generated, then K is finitely generated.

Definition. If K is a module, then a ***subquotient*** of K is a module isomorphic to S/T, where $T \subseteq S \subseteq K$ are submodules.

Thus, a subquotient of K is a quotient of a submodule. It is also easy to see that a subquotient of K is also a submodule of a quotient ($S/T \subseteq K/T$).

Example 10.134.

(i) All the factor modules of a filtration of a module K are subquotients of K.

(ii) The nth homology group of a complex $(\mathbf{C}_\bullet, d_\bullet)$ is a subquotient of C_n. ◂

A spectral sequence computes a homology group H_n in the sense that it computes the factor modules of some filtration of H_n. In general, this gives only partial information about H_n, but, if the factor modules are heavily constrained, then they can give much more information and, indeed, might even determine H_n completely. For example, suppose that only one of the factor modules of K is nonzero, say, $K_i/K_{i+1} \cong A \neq \{0\}$; we claim that $K \cong A$. The beginning of the filtration is

$$K = K_0 \supseteq K_1 \supseteq \cdots \supseteq K_i.$$

Since $K_0/K_1 = \{0\}$, we have $K = K_0 = K_1$. Similarly, $K_1/K_2 = \{0\}$ gives $K_1 = K_2$; indeed, $K = K_0 = K_1 = \cdots = K_i$. Similar reasoning computes the end of the filtration. For example, since $K_{\ell-1}/K_\ell = \{0\}$, we have $K_{\ell-1} = K_\ell = \{0\}$. Thus, the filtration is

$$K = K_0 = \cdots = K_i \supsetneq K_{i+1} = \cdots = K_\ell = \{0\},$$

and so $K \cong K/\{0\} = K_i/K_{i+1} \cong A$.

In order to appreciate spectral sequences, we must recognize an obvious fact: very general statements can become useful if extra simplifying hypotheses can be imposed.

Spectral sequences usually arise in the following context. A ***bigraded module*** $M = M_{\bullet\bullet}$ is a doubly indexed family of modules $M_{p,q}$, where $p, q \in \mathbb{Z}$; we picture a bigraded module as a collection of modules, one sitting on each lattice point (p, q) in the plane.

Thus, there are ***first quadrant*** bigraded modules, for example, with $M_{p,q} = \{0\}$ if either p or q is negative; similarly, there are ***third quadrant*** bigraded modules. A ***bicomplex*** is a bigraded module that has vertical arrows $d''_{p,q} : M_{p,q} \to M_{p,q-1}$ making the columns complexes, horizontal arrows $d'_{p,q} : M_{p,q} \to M_{p-1,q}$ making the rows complexes, and whose squares anticommute:

$$\begin{array}{ccc} M_{p-1,q} & \xleftarrow{\;d'_{p,q}\;} & M_{p,q} \\ {\scriptstyle d''_{p-1,q}}\downarrow & & \downarrow{\scriptstyle d''_{p,q}} \\ M_{p-1,q-1} & \xleftarrow[\;d'_{p,q-1}\;]{} & M_{p,q-1} \end{array}$$

that is, $d'd'' + d''d' = 0$. The reason for the anticommutativity is to allow us to define the ***total complex***, $\text{Tot}(M)$, of a bicomplex M: Its term in degree n is:

$$\text{Tot}(M)_n = \sum_{p+q=n} M_{p,q};$$

its differentiation $d_n : \text{Tot}(M)_n \to \text{Tot}(M)_{n-1}$ is given by

$$d_n = \sum_{p+q=n} d'_{p,q} + d''_{p,q}.$$

Anticommutativity forces $d_{n-1}d_n = 0$:

$$dd = (d' + d'')(d' + d'') = d'd' + (d'd'' + d''d') + d''d'' = 0;$$

thus, $\text{Tot}(M)$ is a complex.

All bigraded modules form a category. Given an ordered pair of integers (a, b), a family of maps $f_{p,q} : M_{p,q} \to L_{p+a,q+b}$ is called a ***map*** $f : M_{\bullet\bullet} \to L_{\bullet\bullet}$ of ***bidegree*** (a, b). For example, the maps d' and d'' above have respective bidegrees $(0, -1)$ and $(-1, 0)$. It is easy to check that all bigraded modules and all maps having some bidegree form a category. One nice feature of composition is that bidegrees add: if f has bidegree (a, b) and f' has bidegree (a', b'), then their composite $f'f$ has bidegree $(a + a', b + b')$. Maps of bigraded modules are used in establishing certain exact sequences. For example, one proof of the five term exact sequence on page 882 uses these maps.

A ***spectral sequence*** is a sequence of bicomplexes, $E^r_{p,q}$, for all $r \geq 2$, where each $E^{r+1}_{p,q}$ is a subquotient of $E^r_{p,q}$ (we must also specify that the homomorphisms of the bicomplex $E^{r+1}_{p,q}$ arise from those of $E^r_{p,q}$). Most spectral sequences arise from a filtration of $\text{Tot}(M)$, where $M_{\bullet\bullet}$ is a bicomplex. In particular, there are two "usual" filtrations (if $M_{\bullet\bullet}$ is either first quadrant or third quadrant), and the spectral sequences they determine are denoted by ${}^{\text{I}}E^r_{p,q}$ and ${}^{\text{II}}E^r_{p,q}$.

We say that a spectral sequence $E^r_{p,q}$ ***converges*** to a (singly graded) module $H_\bullet$, denoted by $E^2_{p,q} \Rightarrow H_n$, if each H_n has a filtration with factor modules

$$E_{0,n}, E_{1,n-1}, \ldots, E_{n,0},$$

and, for all p, q with $p+q=n$, the factor module $E_{p,q}$ is a subquotient of $E^2_{p,q}$. There are two steps to establish before using spectral sequences.

Theorem I. *If $M_{\bullet\bullet}$ is a first quadrant or third quadrant bicomplex, then*

$$ {}^{\mathrm{I}}E^2_{p,q} \Rightarrow H_n(\mathrm{Tot}(M)) \quad \textit{and} \quad {}^{\mathrm{II}}E^2_{p,q} \Rightarrow H_n(\mathrm{Tot}(M)). $$

Thus, for each n, there are two filtrations of $\mathrm{Tot}(M)_n$; one whose factor modules are subquotients of ${}^{\mathrm{I}}E^2_{p,q}$, and another whose factor modules are subquotients of ${}^{\mathrm{II}}E^2_{p,q}$ (as usual, $p+q=n$ in this context), and both converge to the same thing.

Theorem II. *If $M_{\bullet\bullet}$ is a first quadrant or third quadrant bicomplex, then there are formulas for ${}^{\mathrm{I}}E^2_{p,q}$ and ${}^{\mathrm{II}}E^2_{p,q}$ for every p, q.*

Theorem II offers the possibility that subquotients of $E^2_{p,q}$ can be computed.

We illustrate the technique by sketching a proof that $\mathrm{Tor}_n(A, B)$ does not depend on the variable resolved; that is, the value of $\mathrm{Tor}_n(A, B)$, defined as $H_n(\mathbf{P}_A \otimes B)$, where $\mathbf{P}_A$ is a deleted projective resolution of A, coincides with $\mathrm{Tor}_n(A, B)$, defined as $H_n(A \otimes \mathbf{Q}_B)$, where $\mathbf{Q}_B$ is a deleted projective resolution of B. The idea is to resolve both variables simultaneously, using resolutions of each. Define a first quadrant bigraded module $M = \mathbf{P}_A \otimes \mathbf{Q}_B$ whose p, q term is $P_p \otimes Q_q$; make this into a bicomplex by defining vertical arrows $d''_{p,q} = (-1)^p 1 \otimes \partial_q \colon P_p \otimes Q_q \to P_p \otimes Q_{q-1}$ and horizontal arrows $d'_{p,q} = \Delta_p \otimes 1 \colon P_p \otimes Q_q \to P_{p-1} \otimes Q_q$, where the ∂_n are the differentiations in $\mathbf{Q}_B$ and the Δ_n are the differentiations in $\mathbf{P}_A$ (the signs force anticommutativity). The formula whose existence is stated in Theorem II for the first spectral sequence ${}^{\mathrm{I}}E^2_{p,q}$ gives, in this case,

$$ {}^{\mathrm{I}}E^2_{p,q} = \begin{cases} \{0\} & \text{if } q > 0; \\ H_p(\mathbf{P}_A \otimes B) & \text{if } q = 0. \end{cases} $$

Since a subquotient of $\{0\}$ must be $\{0\}$, all but one of the factor modules of a filtration of $H_n(\mathrm{Tot}(M))$ are zero, and so

$$ H_n(\mathrm{Tot}(M)) \cong H_n(\mathbf{P}_A \otimes B). $$

Similarly, the formula alluded to in Theorem II for the second spectral sequence gives

$$ {}^{\mathrm{II}}E^2_{p,q} = \begin{cases} \{0\} & \text{if } p > 0; \\ H_q(A \otimes \mathbf{Q}_B) & \text{if } p = 0. \end{cases} $$

Again, there is a filtration of $H_n(\mathrm{Tot}(M))$ with only one possible nonzero factor module, and so

$$ H_n(\mathrm{Tot}(M)) \cong H_n(A \otimes \mathbf{Q}_B). $$

Therefore,

$$ H_n(\mathbf{P}_A \otimes B) \cong H_n(\mathrm{Tot}(M)) \cong H_n(\mathbf{P}_A \otimes B). $$

We have shown that Tor is independent of the variable resolved.

Here is a cohomology result illustrating how spectral sequences can be used to compute composite functors. The index raising convention extends here, so that one denotes the modules in a third quadrant bicomplex by $M^{p,q}$ instead of by $M_{-p,-q}$.

Theorem 10.135 (Grothendieck). *Let $F : \mathcal{B} \to \mathcal{C}$ and $G : \mathcal{A} \to \mathcal{B}$ be additive functors, where $\mathcal{A}$, $\mathcal{B}$, and $\mathcal{C}$ are module categories. If F is left exact and if E injective in $\mathcal{A}$ implies $(R^m F)(GE) = \{0\}$ for all $m > 0$ (where $R^m F$ are the right derived functors of F), then for every module $A \in \mathcal{A}$, there is a third quadrant spectral sequence*

$$E_2^{p,q} = (R^p F)(R^q G(A)) \Rightarrow R^n(FG)(A).$$

For a proof, see Rotman, *An Introduction to Homological Algebra*, page 350.

The next result shows that if N is a normal subgroup of a group Π, then the cohomology groups of N and of Π/N can be used to compute the cohomology groups of Π.

Theorem 10.136 (Lyndon–Hochschild–Serre). *Let Π be a group with normal subgroup N. For each Π-module A, there is a third quadrant spectral sequence with*

$$E_2^{p,q} = H^p(\Pi/N, H^q(N, A)) \Rightarrow H^n(\Pi, A).$$

Proof. Define functors $G : {}_{\mathbb{Z}\Pi}\mathbf{Mod} \to {}_{\mathbb{Z}(\Pi/N)}\mathbf{Mod}$ and $F : {}_{\mathbb{Z}(\Pi/N)}\mathbf{Mod} \to \mathbf{Ab}$ by $G = \operatorname{Hom}_N(\mathbb{Z}, \ )$ and $F = \operatorname{Hom}_{\Pi/N}(\mathbb{Z}, \ )$. Of course, F is left exact, and it is easy to see that $FG = \operatorname{Hom}_\Pi(\mathbb{Z}, \ )$. A proof that $H^m(\Pi/N, E) = \{0\}$ whenever E is an injective Π-module and $m > 0$ can be found in Rotman, *An Introduction to Homological Algebra*, page 307. The result now follows from Theorem 10.135. •

This theorem was found by Lyndon in his dissertation in 1948, in order to compute the cohomology groups of finitely generated abelian groups Π. Several years later, Hochschild and Serre put the result into its present form.

11

Commutative Rings III

11.1 Local and Global

Quite often, it is easier to examine algebraic structures "one prime at a time." Let G and H be finite groups. If $G \cong H$, then their Sylow p-subgroups are isomorphic for all primes p; studying G and H *locally* means studying their p-subgroups. This local information is not enough to determine whether $G \cong H$; for example, S_3 and $\mathbb{I}_6$ are nonisomorphic groups having isomorphic Sylow subgroups. The *global problem* assumes that the Sylow p-subgroups of groups G and H are isomorphic, for all primes p, and asks what else is necessary to conclude that $G \cong H$. In the case of groups, this leads to the extension problem and cohomology of groups (but even this is inadequate to solve the global problem: for example, S_3 and $\mathbb{I}_6$ have isomorphic Sylow subgroups and the same composition factors). An illustration of the success of this technique is provided by finite abelian groups. The local problem involves primary components (Sylow subgroups), which are direct sums of cyclic groups, and the global problem is solved by the primary decomposition: Every finite abelian group is the direct sum of its primary components. In this case, the local information is sufficient to solve the global problem. The advantage of the local/global approach is that the local problem is simpler than the global and its solution is valuable. We begin this section with another group-theoretic illustration of local and global investigation, after which we will consider localization of commutative rings.

Definition. Let R be a domain with $Q = \mathrm{Frac}(R)$. If M is an R-module, define

$$\mathrm{rank}(M) = \dim_Q(Q \otimes_R M).$$

For example, the rank of an abelian group G is defined as $\dim_{\mathbb{Q}}(\mathbb{Q} \otimes_{\mathbb{Z}} G)$.

Recall that if R is a domain, then an R-module M is *torsion-free* if it has no nonzero elements of finite order; that is, if $r \in R$ and $m \in M$ are nonzero, then rm is nonzero.

Lemma 11.1. *Let R be a domain with $Q = \text{Frac}(R)$ and let M be a torsion-free R-module. Then M has rank 1 if and only if it is isomorphic to a nonzero R-submodule of Q.*

Proof. If $\text{rank}(M) = 1$, then $M \neq \{0\}$. Exactness of $0 \to R \to Q$ gives exactness of

$$\text{Tor}_1^R(Q/R, M) \to R \otimes_R M \to Q \otimes_R M.$$

By Lemma 10.101(iii), we have $\text{Tor}_1^R(Q/R, M) \cong tM$, the torsion submodule of M, and so $\text{Tor}_1^R(Q/R, M) = \{0\}$ because M is torsion-free. But Proposition 8.86 gives $R \otimes_R M \cong M$, while $Q \otimes_R M \cong Q$ because M has rank 1. Therefore, M is isomorphic to an R-submodule of Q.

Conversely, if M is isomorphic to an R-submodule of Q, there is an exact sequence $0 \to M \to Q$. Since Q is a flat R-module, by Corollary 8.103, we have exactness of $0 \to Q \otimes_R M \to Q \otimes_R Q$. This is an exact sequence of vector spaces over Q, with $Q \otimes_R Q \cong Q$ being one-dimensional. Therefore, the nonzero subspace $Q \otimes_R M$ is also one-dimensional; that is, $\text{rank}(M) = 1$. •

Example 11.2.
The following abelian groups are torsion-free of rank 1:

(i) The group $\mathbb{Z}$ of integers;

(ii) The additive group $\mathbb{Q}$;

(iii) the set of all rationals having a finite decimal expansion;

(iv) the set of all rationals having squarefree denominator. ◂

Proposition 11.3. *Let R be a domain with $Q = \text{Frac}(R)$. Two submodules A and B of Q are isomorphic if and only if there is $c \in Q$ with $B = cA$.*

Proof. If $B = cA$, then $A \cong B$ via $a \mapsto ca$.

Conversely, suppose that $f : A \to B$ is an isomorphism. We show first that if $a \in A$ is nonzero, then f is determined by its values on $\langle a \rangle$: If $g : A \to B$ and $g|\langle a \rangle = f|\langle a \rangle$, then $f = g$. If $x \in A$, then there are $r, s \in R$ with $sx = ra \in \langle a \rangle$ (because A is a submodule of Q), and so

$$f(sx) = f(ra) = rf(a) = rg(a) = g(ra) = g(sx).$$

Hence, $s(f(x) - g(x)) = 0$, and, since B is torsion-free, we have $f(x) = g(x)$.

If $f(a) = b$, define $c = b/a$. In order to show that $f(x) = cx$ for all $x \in A$, it now suffices to prove that $f(ra) = c(ra)$ for all $r \in R$. But

$$f(ra) = rf(a) = rb = r(b/a)a = c(ra).$$

It follows that $f(x) = cx$ for all $x \in A$ and that $B = cA$. •

Definition. For each prime p, we define a subring of $\mathbb{Q}$,

$$\mathbb{Z}_{(p)} = \{a/b \in \mathbb{Q} : (b, p) = 1\}.$$

Proposition 11.4.

(i) *For each prime p, the ring $\mathbb{Z}_{(p)}$ is a local*[1] PID.

(ii) *If G is a torsion-free abelian group of rank 1, then $\mathbb{Z}_{(p)} \otimes_{\mathbb{Z}} G$ is a torsion-free $\mathbb{Z}_{(p)}$-module of rank 1.*

(iii) *If M is a torsion-free $\mathbb{Z}_{(p)}$-module of rank 1, then $M \cong \mathbb{Z}_{(p)}$ or $M \cong \mathbb{Q}$.*

Proof. (i) We show that the only nonzero ideals I in $\mathbb{Z}_{(p)}$ are (p^n), for $n \geq 0$; it will then follow that $\mathbb{Z}_{(p)}$ is a PID and that (p) is its unique maximal ideal. Since $\mathbb{Z}_{(p)} \subseteq \mathbb{Q}$, each nonzero $x \in p^{-1}\mathbb{Z}$ has the form a/b for integers a and b, where $(b, p) = 1$. But $a = p^n a'$, where $n \geq 0$ and $(a', p) = 1$; that is, there is a unit $u \in \mathbb{Z}_{(p)}$, namely, $u = a'/b$, with $x = up^n$. Let $I \neq \{0\}$ be an ideal. Of all the nonzero elements in I, choose $x = up^n \in I$, where u is a unit, with n minimal. Then $I = (x) = (p^n)$, for if $y \in I$, then $y = vp^m$, where v is a unit and $n \leq m$. Hence, $p^n \mid y$ and $y \in (p^n)$.

(ii) Since $\mathbb{Z}_{(p)} \subseteq \mathbb{Q}$, it is an additive torsion-free abelian group of rank 1, and so it is flat (Corollary 9.6). Hence, exactness of $0 \to G \to \mathbb{Q}$ gives exactness of

$$0 \to \mathbb{Z}_{(p)} \otimes_{\mathbb{Z}} G \to \mathbb{Z}_{(p)} \otimes_{\mathbb{Z}} \mathbb{Q}.$$

By Exercise 11.5 on page 920, $\mathbb{Z}_{(p)} \otimes_{\mathbb{Z}} \mathbb{Q} \cong \mathbb{Q} = \mathrm{Frac}(\mathbb{Z}_{(p)})$, so that $\mathbb{Z}_{(p)} \otimes_{\mathbb{Z}} G$ is a torsion-free $\mathbb{Z}_{(p)}$-module of rank 1.

(iii) There is no loss in generality in assuming that $M \subseteq \mathbb{Q}$ and that $1 \in M$. Consider the equations $p^n y_n = 1$ for $n \geq 0$. We claim that if all these equations are solvable for $y_n \in M$, then $M = \mathbb{Q}$. If $a/b \in \mathbb{Q}$, then $a/b = a/p^n b'$, where $(b', p) = 1$, and so $a/b = (a/b')y_n$; as $a/b' \in \mathbb{Z}_{(p)}$, we have $a/b \in M$. We may now assume that there is a largest $n \geq 0$ for which the equation $p^n y_n = 1$ is solvable for $y_n \in M$. We claim that $M = \langle y_n \rangle$, the cyclic submodule generated by y_n, which will show that $M \cong \mathbb{Z}_{(p)}$. If $m \in M$, then $m = c/d = p^r c'/p^s d' = (c'/d')(1/p^{s-r})$, where $(c', p) = 1 = (d', p)$. Since c'/d' is a unit in $\mathbb{Z}_{(p)}$, we have $1/p^{s-r} \in M$, and so $s - r \leq n$; that is, $s - r = n - \ell$ for some $\ell \geq 0$. Hence, $1/p^{s-r} = 1/p^{n-\ell} = p^\ell/p^n = p^\ell y_n$, and so $m = (c'p^\ell/d')y_n \in \langle y_n \rangle$. •

Definition. A ***discrete valuation ring***, abbreviated DVR, is a local PID that is not a field.

For example, $\mathbb{Z}_{(p)}$ is a DVR.

---

[1]Recall that a *local ring* is a commutative ring having a unique maximal ideal. Most authors insist that local rings are noetherian [$\mathbb{Z}_{(p)}$ is even a PID]. Other authors allow local rings to be noncommutative, defining a ring R to be local if it has a unique maximal left ideal $\mathfrak{m}$. In this case, $\mathfrak{m} = J(R)$, the Jacobson radical, so that it is a two-sided ideal.

Definition. Two torsion-free abelian groups of rank 1, G and H, are ***locally isomorphic*** if $\mathbb{Z}_{(p)} \otimes_{\mathbb{Z}} G \cong \mathbb{Z}_{(p)} \otimes_{\mathbb{Z}} H$ for all primes p.

We have solved the local problem for torsion-free abelian groups G of rank 1; associate to G the family $\mathbb{Z}_{(p)} \otimes_{\mathbb{Z}} G$ of $\mathbb{Z}_{(p)}$-modules, one for each prime p.

Example 11.5.
Let G be the subgroup of $\mathbb{Q}$ consisting of those rationals having squarefree denominator. Then G and $\mathbb{Z}$ are locally isomorphic, but they are not isomorphic, because G is not finitely generated. ◂

We now consider the global problem for torsion-free abelian groups of rank 1.

Definition. Let G be an abelian group. If $x \in G$ and p is a prime, we say that x is ***divisible by*** p^n ***in*** G if there exists $y_n \in G$ with $p^n y_n = x$. Define the p-***height*** of x, denoted by $h_p(x)$, by

$$h_p(x) = \begin{cases} \infty & \text{if } x \text{ is divisible by } p^n \text{ in } G \text{ for all } n \geq 0 \\ k & \text{if } x \text{ is divisible by } p^k \text{ in } G \text{ but not by } p^{k+1}. \end{cases}$$

The ***height sequence*** (or ***characteristic***) of x in G, where x is nonzero, is the sequence

$$\chi(x) = \chi_G(x) = (h_2(x), h_3(x), h_5(x), \ldots, h_p(x), \ldots).$$

Thus, $\chi(x)$ is a sequence (h_p), where $h_p = \infty$ or $h_p \in \mathbb{N}$. Let $G \subseteq \mathbb{Q}$ and let $x \in G$ be nonzero. If $\chi(x) = (h_p)$ and $a = p_1^{f_1} \cdots p_n^{f_n}$, then $\frac{1}{a}x \in G$ if and only if $f_{p_i} \leq h_{p_i}$ for $i = 1, \ldots, n$.

Example 11.6.
Each of the groups in Example 11.2 contains $x = 1$.
(i) In $\mathbb{Z}$,

$$\chi_{\mathbb{Z}}(1) = (0, 0, 0, \ldots).$$

(ii) In $\mathbb{Q}$,

$$\chi_{\mathbb{Q}}(1) = (\infty, \infty, \infty, \ldots).$$

(iii) If G is the group of all rationals having a finite decimal expansion, then

$$\chi_G(1) = (\infty, 0, \infty, 0, 0, \ldots).$$

(iv) If H is the group of rationals having squarefree denominators, then

$$\chi_H(1) = (1, 1, 1, \ldots). \quad ◂$$

Different elements in a torsion-free abelian group of rank 1 may have different height sequences. For example, if G is the group of rationals having finite decimal expansions, then 1 and $\frac{63}{8}$ lie in G, and

$$\chi(1) = (\infty, 0, \infty, \ldots) \quad \text{and} \quad \chi(\tfrac{63}{8}) = (\infty, 2, \infty, 1, 0, 0, \ldots).$$

Thus, these height sequences agree for infinite p-heights, but they disagree for two finite p-heights.

Definition. Two height sequences $(h_2, h_3, \ldots, h_p, \ldots)$ and $(k_2, k_3, \ldots, k_p, \ldots)$ are ***equivalent***, denoted by

$$(h_2, h_3, \ldots, h_p, \ldots) \sim (k_2, k_3, \ldots, k_p, \ldots),$$

if there are only finitely many p for which $h_p \neq k_p$ and, for such primes p, neither h_p nor k_p is ∞.

It is routine to see that equivalence is, in fact, an equivalence relation.

Lemma 11.7. *If G is a torsion-free abelian group of rank 1, and if $x, y \in G$ are nonzero, then their height sequences $\chi(x)$ and $\chi(y)$ are equivalent.*

Proof. We may assume that $G \subseteq \mathbb{Q}$. If $b = p_1^{e_1} \cdots p_n^{e_n}$, then it is easy to see that $h_p(bx) = h_p(x)$ for all $p \notin \{p_1, \ldots, p_n\}$, while

$$h_{p_i}(bx) = e_i + h_{p_i}(x)$$

for $i = 1, \ldots, n$ (we agree that $e_i + \infty = \infty$). Hence, $\chi(x) \sim \chi(bx)$. Since $x, y \in G \subseteq \mathbb{Q}$, we have $x/y = a/b$ for integers a, b, so that $bx = ay$. Therefore, $\chi(x) \sim \chi(bx) = \chi(ay) \sim \chi(y)$. •

Definition. The equivalence class of a height sequence is called a ***type***. If G is a torsion-free abelian group of rank 1, then its ***type***, denoted by $\tau(G)$, is the type of a height sequence $\chi(x)$, where x is a nonzero element of G.

Lemma 11.7 shows that $\tau(G)$ depends only on G and not on the choice of nonzero element $x \in G$. We now solve the global problem.

Theorem 11.8. *If G and H are torsion-free abelian groups of rank 1, then $G \cong H$ if and only if $\tau(G) = \tau(H)$.*

Proof. Let $\varphi\colon G \to H$ be an isomorphism. If $x \in G$ is nonzero, it is easy to see that $\chi(x) = \chi(\varphi(x))$, and so $\tau(G) = \tau(H)$.

For the converse, there is no loss in generality in assuming that both G and H are subgroups of $\mathbb{Q}$. Choose nonzero $x \in G$ and $y \in H$. By the definition of equivalence, there are primes $p_1, \ldots, p_n, q_1, \ldots, q_m$ with $h_{p_i}(x) < h_{p_i}(y) < \infty$, with $\infty > h_{q_j}(x) > h_{q_j}(y)$, and with $h_p(x) = h_p(y)$ for all other primes p. Define $b = \prod p_i^{h_{p_i}(y) - h_{p_i}(x)}$. Then

$bx \in G$ and $h_{p_i}(bx) = \big(h_{p_i}(y) - h_{p_i}(x)\big) + h_{p_i}(x) = h_{p_i}(y)$. A similar construction, using $a = \prod q_j^{h_{q_j}(x) - h_{q_j}(y)}$, gives $\chi(bx) = \chi(ay)$. We have found elements $x' = bx \in G$ and $y' = ay \in H$ having the same height sequence.

Define $\varphi\colon G \to \mathbb{Q}$ by $\varphi(g) = \frac{y'}{x'}g$. It is obvious that φ is an injective homomorphism. We claim that $\operatorname{im}\varphi \subseteq H$. Since every $g \in G$ can be written as $g = \frac{1}{c}x'$, it suffices to show that if $\frac{1}{c}x' \in G$, then $\varphi(\frac{1}{c}x') = \frac{1}{c}y' \in H$. But if $c = p_1^{f_1} \cdots p_t^{f_t}$, then $\frac{1}{c}x' \in G$ if and only if $f_p \le h_p(x')$. Since $\chi(x') = \chi(y')$, $\frac{1}{c}y' \in H$ if and only if $f_p \le h_p(y') = h_p(x')$. Thus, we may view φ as a map $G \to H$. Finally, to see that φ is a surjection, note that its inverse is given by $h \mapsto \frac{x'}{y'}h$, which is a map $H \to G$. •

The uniqueness theorem just proved is complemented by an existence theorem.

Proposition 11.9. *Given a height sequence* $(k_2, k_3, \ldots, k_p, \ldots)$, *where* $0 \le k_p \le \infty$, *there exists a unique subgroup* $G \subseteq \mathbb{Q}$ *containing* 1 *with* $h_p(1) = k_p$ *for all* p. *Thus, given any type* τ, *there exists a torsion-free abelian group* G *of rank* 1, *unique to isomorphism, with* $\tau(G) = \tau$.

Proof. Define

$$D = \{a \in \mathbb{Z} : a = \prod p_i^{e_i} \text{ with } 0 \le e_i \le k_{p_i} \text{ for all } i\}$$

(if $k_{p_i} = \infty$, then $0 \le e_i \le k_{p_i}$ means that $e_i \in \mathbb{N}$), and define

$$G = \{m/a \in \mathbb{Q} : m \in \mathbb{Z} \text{ and } a \in D\}.$$

To see that G is a subgroup of $\mathbb{Q}$, it suffices to prove that it is closed under addition. Let m/a and n/b be in G, where $a = \prod p_i^{e_i}$, $b = \prod p_i^{f_i}$, and $\max\{e_i, f_i\} \le k_{p_i}$; that is, $[a, b] \in D$. Now

$$\frac{m}{a} + \frac{n}{b} = \frac{ma' + nb'}{[a, b]},$$

where $[a, b] = \operatorname{lcm}\{a, b\} = \prod p_i^{\max\{e_i, f_i\}}$, $a' = [a, b]/a$, and $b' = [a, b]/b$. Since $[a, b] \in D$, we have $m/a + n/b \in G$. It is clear that $h_p(1) = k_p$ for all p, and so $\tau(G) = \tau$.

Let us now prove uniqueness. Let G and H be subgroups of $\mathbb{Q}$ containing 1 with $\chi_H(1) = \chi_G(1)$. Suppose that $m/d \in H$ is in lowest terms—$(m, d) = 1$. Then there are integers s and t with $1 = sm + td$, so that $1/d = s(m/d) + td/d \in H$. On the other hand, $1/d \in G$, by definition of height sequence. It follows that $H \subseteq G$, for H is generated by all elements of the form $1/d$. The reverse inclusion is proved similarly, and so $G = H$. •

Corollary 11.10.

(i) *There are uncountably many nonisomorphic subgroups of* $\mathbb{Q}$.

(ii) *If* R *is a subring of* $\mathbb{Q}$, *then the height sequence of* 1 *consists of* 0*'s and* ∞*'s.*

(iii) *There are uncountably many nonisomorphic subrings of* $\mathbb{Q}$. *In fact, distinct subrings of* $\mathbb{Q}$ *are not isomorphic as rings.*

Proof. (i) Given any type τ, Proposition 11.9 provides a torsion-free abelian group G of rank 1 with $\tau(G) = \tau$. But there are uncountably many types; for example, two height sequences of 0's and ∞'s are equivalent if and only if they are equal.

(ii) If $h_p(1) > 0$, then $\frac{1}{p} \in R$. Since R is a ring, $\left(\frac{1}{p}\right)^n = \frac{1}{p^n} \in R$ for all $n \geq 1$, and so $h_p(1) = \infty$.

(iii) If R and S are distinct subrings of $\mathbb{Q}$, then the height sequences of 1 are distinct, by part (ii). Both statements follow from the observation that two height sequences whose only terms are 0 and ∞ are equivalent if and only if they are equal. •

A. G. Kurosh classified torsion-free abelian groups G of finite rank n with invariants $n = \text{rank}(G)$, $\dim(\mathbb{F}_p \otimes G)$ for all primes p, and an equivalence class of sequences (M_p), where M_p is an $n \times n$ nonsingular matrix over the p-adic numbers $\mathbb{Q}_p$ (this theorem is not easy to use, for it is almost impossible to determine whether two groups have equivalent matrix sequences). It is easy to see that every such group G is a direct sum of indecomposable[2] groups; however, there is virtually no uniqueness for such a decompostion. For example, there exists a group G with

$$G = A_1 \oplus A_2 = B_1 \oplus B_2 \oplus B_3,$$

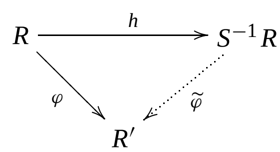
with all the summands indecomposable, with $\text{rank}(A_1) = 1$, $\text{rank}(A_2) = 5$, and with $\text{rank}(B_j) = 2$ for $j = 1, 2, 3$. Thus, the number of indecomposable summands in a decomposition is not uniquely determined by G, nor is the isomorphism class of any of the indecomposable summands. Here is an interesting theorem of A. L. S. Corner (that can be used to produce bad examples of torsion-free groups such as the group G just discussed). Let R be a ring whose additive group is countable, torsion-free, and reduced (it has no nonzero divisible subgroups). Then there exists an abelian group G, also countable, torsion-free, and reduced, with $\text{End}(G) \cong R$. Moreover, if the additive group of R has finite rank n, then G can be chosen to have rank $2n$. For a proof, see Fuchs, *Infinite Abelian Groups* II, page 231.

The local approach to commutative rings generalizes the construction of the local rings $\mathbb{Z}_{(p)}$ from $\mathbb{Z}$. Given a subset S of a commutative ring R closed under multiplication, most authors construct the localization $S^{-1}R$ by generalizing the (tedious) construction of the fraction field of a domain R. They define a relation on $R \times S$ by $(r, \sigma) \equiv (r', \sigma')$ if there exists $\sigma'' \in S$ with $\sigma''(r\sigma' - r'\sigma) = 0$ (this definition reduces to the usual definition involving cross multiplication when R is a domain and S is the subset of all nonzero elements). After proving that this is an equivalence relation, $S^{-1}R$ is defined to be the set of all equivalence classes, addition and multiplication are defined and proved to be well-defined, all the R-algebra axioms are verified, and the elements of S are shown to be invertible. In other

[2] An abelian group G is *indecomposable* if there do not exist nonzero groups A and B with $G \cong A \oplus B$.

words, we regard the elements of $S^{-1}R$ as fractions with denominators in S. We prefer to develop the existence and first properties of $S^{-1}R$ in another manner, which is less tedious and which will show how the equivalence relation generalizing cross multiplication arises.

Definition. Let R be a commutative ring and let S be any subset of R. A ***localization*** of R is an R-algebra $S^{-1}R$ and an R-algebra map $h\colon R \to S^{-1}R$, called the ***localization map***, such that $h(s)$ is invertible in $S^{-1}R$, for every $s \in S$, and $S^{-1}R$ is a solution to the following universal mapping problem.



If R' is a commutative R-algebra and $\varphi\colon R \to R'$ is an R-algebra map for which $\varphi(s)$ is invertible in R' for all $s \in S$, then there exists a unique R-algebra map $\widetilde{\varphi}\colon S^{-1}R \to R'$ with $\widetilde{\varphi}h = \varphi$.

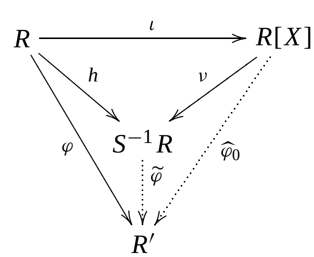
The localization $S^{-1}R$, as any solution to a universal mapping problem, is unique up to isomorphism if it exists.

Theorem 11.11. *For every subset S of a commutative ring R, the localization $S^{-1}R$ exists.*

Proof. Let $X = \{x_s : s \in S\}$ be a set with $x_s \mapsto s$ a bijection $X \to S$, and let $R[X]$ be the polynomial ring over R with variables X. Define

$$S^{-1}R = R[X]/I,$$

where I is the ideal generated by $\{sx_s - 1 : s \in S\}$, and define $h\colon R \to S^{-1}R$ by $h\colon r \mapsto r + I$, where r is a constant polynomial. It is clear that $S^{-1}R$ is an R-algebra, that h is an R-algebra map, and that each $h(s)$ is invertible. Assume now that R' is an R-algebra, and that $\varphi\colon R \to R'$ is an R-algebra map with $\varphi(s)$ invertible for all $s \in S$. Consider the diagram in which the top arrow $\iota\colon R \to R[X]$ sends each $r \in R$ to the constant polynomial r and $\nu\colon R[X] \to S^{-1}R$ is the natural map.



The top triangle commutes because both h and $\nu\iota$ send $r \in R$ to $r + I$. Since $R[X]$ is the free commutative R-algebra on X, there is an R-algebra map $\varphi_0\colon R[X] \to R'$ with

$\varphi_0(x_s) = \varphi(s)^{-1}$ for all $s \in S$. Clearly, $I \subseteq \ker \varphi_0$, for $\varphi_0(sx_s - 1) = 0$, and so there is an R-algebra map $\widetilde{\varphi}\colon S^{-1}R = R[X]/I \to R'$ making the diagram commute. That $\widetilde{\varphi}$ is the unique such map follows from $S^{-1}R$ being generated by $\operatorname{im} h \cup \{h(s)^{-1} : s \in S\}$ as an R-algebra. •

The next definition is natural in this context, for if s, s' are invertible elements in some commutative ring, then their product ss' is also invertible.

Definition. A subset S of a commutative ring R is ***multiplicatively closed*** if $1 \in S$ and $s, s' \in S$ implies $ss' \in S$. Every commutative ring is a multiplicative monoid. If S is any (possibly empty) subset of R, then

$$\overline{S} = \text{ the submonoid of } R \text{ generated by } S.$$

We call $\overline{S}$ the multiplicatively closed subset ***generated*** by S.

Exercise 11.8 on page 920 says that $(\overline{S})^{-1}R \cong S^{-1}R$.

Example 11.12.
(i) If R is a commutative ring and $s \in R$, then $\overline{\{s\}} = \{1, s, s^2, s^3, \ldots\}$ is the multiplicatively closed set generated by the element s.

(ii) If $\mathfrak{p}$ is a prime ideal in R, then $a \notin \mathfrak{p}$ and $b \notin \mathfrak{p}$ imply $ab \notin \mathfrak{p}$. In other words, the complement $R - \mathfrak{p}$ is multiplicatively closed.

(iii) Let P be the set of all primes in $\mathbb{Z}$. If $S \subseteq P$, then

$$\overline{S} = \left\{ p_1^{e_1} \cdots p_n^{e_n} : p_i \in S \text{ and } e_i \geq 0 \right\}. \quad ◀$$

We now describe the elements in $S^{-1}R$.

Proposition 11.13. *If S is a subset of a commutative ring R, then each $y \in S^{-1}R$ has a* (*not necessarily unique*) *factorization*

$$y = h(r)h(\sigma)^{-1}, \quad \textit{where } r \in R \textit{ and } \sigma \in \overline{S}.$$

Proof. The existence theorem constructs $S^{-1}R$ as $R[X]/I$, where $X = \{x_s : s \in S\}$ and $I = \{sx_s - 1 : s \in S\}$. Thus, each $y \in S^{-1}R$ has the form $y = f(x_1, \ldots, x_n) + I$, where $x_i = x_{s_i}$ for some $s_i \in S$. The proposition is proved by induction on $n \geq 0$. If $n = 0$, then $f \in R$ and $y = h(f)$. For the inductive step, let $y = f(x_1, \ldots, x_n) + I$. Write $(x_1, \ldots, x_{n-1}) = X$, $x_n = x$, and

$$f(X, x_n) = g_0(X) + g_1(X)x + \cdots + g_m(X)x^m,$$

where $g_i(X) \in R[X]$. In $S^{-1}R$, we have $x = h(s)^{-1}$ for some $s \in S$ and, by induction, $g_i(X) = h(r_i)h(\sigma_i)^{-1}$, where $r_i \in R$ and $\sigma_i \in \overline{S}$. Therefore,

$$\begin{aligned} y &= h(r_0)h(\sigma_0)^{-1} + h(r_1)h(\sigma_1)^{-1}h(s)^{-1} + \cdots + h(r_m)h(\sigma_m)^{-1}h(s)^{-m} \\ &= h(s)^{-m}\big(h(r_0)h(\sigma_0)^{-1}h(s)^m + h(r_1)h(\sigma_1)^{-1}h(s)^{m-1} + \cdots + h(r_m)h(\sigma_m)^{-1}\big) \\ &= h(r')h(\sigma)^{-1}, \end{aligned}$$

where $r' \in R$ and $\sigma = \sigma_0\sigma_1 \cdots \sigma_m s^m \in \overline{S}$. Therefore, $y = h(r')h(\sigma)^{-1}$. •

In light of Proposition 11.13, the elements of $S^{-1}R$ can be regarded as "fractions" $h(r)h(\sigma)^{-1}$, where $r \in R$ and $\sigma \in \overline{S}$.

Notation. Let $h\colon R \to S^{-1}R$ be the localization map. If $r \in R$ and $\sigma \in \overline{S}$, define

$$r/\sigma = h(r)h(\sigma)^{-1}.$$

In particular, $r/1 = h(r)$.

Is the localization map $h\colon r \mapsto r/1$ an injection? The easiest example in which h has a kernel occurs if $0 \in S$ (after all, S is allowed to be any subset of R). If 0 is invertible, then $0 = 00^{-1} = 1$, and so $S^{-1}R$ is the zero ring. Thus, $h\colon R \to S^{-1}R$ is the zero map, and hence it is not injective unless R is the zero ring. The next lemma investigates $\ker h$.

Proposition 11.14. *If S is a subset of a commutative ring R, and if $h\colon R \to S^{-1}R$ is the localization map, then*

$$\ker h = \{r \in R : \sigma r = 0 \text{ for some } \sigma \in \overline{S}\}.$$

Proof. If $\sigma r = 0$, then $0 = h(\sigma)h(r)$ in $S^{-1}R$. Since $h(\sigma)$ is a unit, we have $0 = h(\sigma)^{-1}h(\sigma)h(r) = h(r)$, and so $r \in \ker h$.

Conversely, suppose that $h(r) = 0$ in $S^{-1}R$. Since $S^{-1}R = R[X]/I$, where $I = (sx_s - 1 : s \in S)$, there is an equation $r = \sum_{i=1}^n f_i(X)(s_i x_{s_i} - 1)$ in $R[X]$. If $S_0 = \{s_1, \ldots, s_n\} \cup \{\text{nonzero coefficients of all } f_i(X)\}$ and $h_0\colon R \to (S_0)^{-1}R$ is the localization map, then $r \in \ker h_0$. In fact, if $s = s_1 \cdots s_n$ and $h'\colon R \to \{s\}^{-1}R$ is the localization map, then every $h'(s_i)$ is invertible, for $s_i^{-1} = s^{-1}s_1 \cdots \widehat{s_i} \cdots s_n$. Now $\{s\}^{-1}R = R[x]/(sx - 1)$, so that $r \in \ker h'$ says that there is $f(x) = \sum_{i=0}^m a_i x^i$ with

$$r = f(x)(sx-1) = \Big(\sum_{i=0}^m a_i x^i\Big)(sx-1) = \sum_{i=0}^m \big(sa_i x^{i+1} - a_i x^i\big) \text{ in } R[x].$$

Expanding and equating coefficients of like powers of x gives

$$r = -a_0, \quad sa_0 = a_1, \quad \ldots, \quad sa_{m-1} = a_m, \quad sa_m = 0.$$

Hence, $sr = -sa_0 = -a_1$, and, by induction, $s^i r = -a_i$ for all i. In particular, $s^m r = -a_m$, and so $s^{m+1}r = -sa_m = 0$, as desired. •

When are two 'fractions" r/σ and r'/σ' equal?

Corollary 11.15. *Let S be a subset of a commutative ring R. If $r/\sigma, r'/\sigma' \in S^{-1}R$, where $\sigma, \sigma' \in \overline{S}$, then $r/\sigma = r'/\sigma'$ if and only if there exists $\sigma'' \in \overline{S}$ with $\sigma''(r\sigma' - r'\sigma) = 0$ in R.*

Remark. If S contains no zero divisors, then $\sigma''(r\sigma' - r'\sigma) = 0$ if and only if $r\sigma' - r'\sigma = 0$, because σ'' is a unit, and so $r\sigma' = r'\sigma$. ◀

Proof. If $r/\sigma = r'/\sigma'$, then multiplying by $\sigma\sigma'$ gives $(r\sigma' - r'\sigma)/1 = 0$ in $S^{-1}R$. Hence, $r\sigma' - r'\sigma \in \ker h$, and Proposition 11.14 gives $\sigma'' \in \overline{S}$ with $\sigma''(r\sigma' - r'\sigma) = 0$ in R.

Conversely, if $\sigma''(r\sigma' - r'\sigma) = 0$ in R for some $\sigma'' \in \overline{S}$, then $h(\sigma'')h(r\sigma' - r'\sigma) = 0$ in $S^{-1}R$. As $h(\sigma'')$ is a unit, we have $h(r)h(\sigma') = h(r')h(\sigma)$; as $h(\sigma)$ and $h(\sigma')$ are units, $h(r)h(\sigma)^{-1} = h(r')h(\sigma')^{-1}$; that is, $r/\sigma = r'/\sigma'$. •

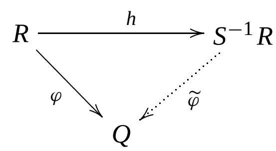
Corollary 11.16. *Let S be a subset of a commutative ring R.*

(i) *If S contains no zero divisors, then the localization map $h\colon R \to S^{-1}R$ is an injection.*

(ii) *If R is a domain with $Q = \mathrm{Frac}(R)$, then $S^{-1}R \subseteq Q$. Moreover, if $S = R - \{0\}$, then $S^{-1}R = Q$.*

Proof. (i) This follows easily from Proposition 11.14.

(ii) The localization map $h\colon R \to S^{-1}R$ is an injection, by Proposition 11.14. Consider the diagram



where φ is the inclusion. If $\widetilde{\varphi}(h(r)h(\sigma)^{-1}) = 0$, then $\widetilde{\varphi}(h(r)) = 0$, because $h(\sigma)$ is a unit in $S^{-1}R$. But commutativity of the diagram gives $\widetilde{\varphi}h(r) = \varphi(r)$. As φ is an injection, $r = 0$; hence, $h(r)h(\sigma)^{-1} = 0$, and so $\widetilde{\varphi}$ is an injection. •

As a consequence of Corollary 11.16(ii), when R is a domain and S is a multiplicatively closed subset not containing 0, then $S^{-1}R$ consists of all elements $a/s \in \mathrm{Frac}(R)$ with $a \in R$ and $s \in S$.

Let us now investigate the ideals in $S^{-1}R$.

Definition. If S is a subset of a commutative ring R, and if I is an ideal in R, then we denote the ideal in $S^{-1}R$ generated by $h(I)$ by $S^{-1}I$.

Example 11.17.

(i) If S is a subset of a commutative ring R, and if I is an ideal in R containing an element $\sigma \in \overline{S}$—that is, $I \cap \overline{S} \neq \varnothing$, then $S^{-1}I$ contains $\sigma/\sigma = 1$, and so $S^{-1}I = S^{-1}R$.

(ii) Let S consist of all the odd integers [that is, S is the complement of the prime ideal (2)], let $I = (3)$, and let $I' = (5)$. Then $S^{-1}I = S^{-1}\mathbb{Z} = S^{-1}I'$. Therefore, the function from the ideals in $\mathbb{Z}$ to the ideals in $S^{-1}\mathbb{Z} = \mathbb{Z}_{(2)}$, given by $I \mapsto S^{-1}I$, is not injective.

In the next corollary, we will see an improvement when we restrict our attention to prime ideals contained in (2). ◀

Corollary 11.18. *Let S be a subset of a commutative ring R.*

(i) *Every ideal J in $S^{-1}R$ is of the form $S^{-1}I$ for some ideal I in R. In fact, if R is a domain and $I = J \cap R$, then $J = S^{-1}I$; in the general case, if $I = h^{-1}(h(R) \cap J)$, then $J = S^{-1}I$.*

(ii) *If I is an ideal in R, then $S^{-1}I = S^{-1}R$ if and only if $I \cap \overline{S} \neq \varnothing$.*

(iii) *If $\mathfrak{q}$ is a prime ideal in R with $\mathfrak{q} \cap \overline{S} = \varnothing$, then $S^{-1}\mathfrak{q}$ is a prime ideal in $S^{-1}R$.*

(iv) *The function $\mathfrak{q} \mapsto S^{-1}\mathfrak{q}$ is a bijection from the family of all prime ideals in R that are disjoint from $\overline{S}$ to* $\operatorname{Spec}(S^{-1}R)$.

(v) *If R is noetherian, then $S^{-1}R$ is also noetherian.*

Proof. (i) Let $J = (j_\lambda : \lambda \in \Lambda)$. By Proposition 11.14, we have $j_\lambda = h(r_\lambda)h(\sigma_\lambda)^{-1}$, where $r_\lambda \in R$ and $\sigma_\lambda \in \overline{S}$. Define I to be the ideal in R generated by $\{r_\lambda : \lambda \in \Lambda\}$; that is, $I = h^{-1}(h(R) \cap J)$. It is clear that $S^{-1}I = J$; in fact, since all σ_λ are units in $S^{-1}R$, we have $J = (h(r_\lambda) : \lambda \in \Lambda)$.

(ii) If $\sigma \in I \cap \overline{S}$, then $\sigma/1 \in S^{-1}I$. But $\sigma/1$ is a unit in $S^{-1}R$, and so $S^{-1}I = S^{-1}R$. Conversely, if $S^{-1}I = S^{-1}R$, then $h(a)h(\sigma)^{-1} = 1$ for some $a \in I$ and $\sigma \in \overline{S}$. Therefore, $\sigma - a \in \ker h$, and so there is $\sigma'' \in \overline{S}$ with $\sigma''(\sigma - a) = 0$. Therefore, $\sigma''\sigma = \sigma''a \in I$. Since $\overline{S}$ is multiplicatively closed, $\sigma''\sigma \in I \cap \overline{S}$.

(iii) Suppose that $\mathfrak{q}$ is a prime ideal in R. First, $S^{-1}\mathfrak{q}$ is a proper ideal, for $\mathfrak{q} \cap \overline{S} = \varnothing$. If $(a/\sigma)(b/\tau) = \mathfrak{q}/\omega$, where $a, b \in R$ and $\sigma, \tau, \omega \in \overline{S}$, then there is $\sigma'' \in \overline{S}$ with $\sigma''(\omega ab - \sigma\tau\mathfrak{q}) = 0$. Hence, $\sigma''\omega ab \in \mathfrak{q}$. Now $\sigma''\omega \notin \mathfrak{q}$ (because $\sigma''\omega \in \overline{S}$ and $\overline{S} \cap \mathfrak{q} = \varnothing$); hence, $ab \in \mathfrak{q}$ (because $\mathfrak{q}$ is prime). Thus, either a or b lies in $\mathfrak{q}$, and either a/σ or b/τ lies in $S^{-1}\mathfrak{q}$. Therefore, $S^{-1}\mathfrak{q}$ is a prime ideal.

(iv) Suppose that $\mathfrak{p}$ and $\mathfrak{q}$ are prime ideals in R with $S^{-1}\mathfrak{p} = S^{-1}\mathfrak{q}$; we may assume that $\mathfrak{p} \cap \overline{S} = \varnothing = \mathfrak{q} \cap \overline{S}$. If $a \in \mathfrak{p}$, then there is $b \in \mathfrak{q}$ and $\sigma \in \overline{S}$ with $a/1 = b/\sigma$. Hence, $\sigma a - b \in \ker h$, where h is the localization map, and so there is $\sigma' \in \overline{S}$ with $\sigma'\sigma a = \sigma' b \in \mathfrak{q}$. But $\sigma'\sigma \in \overline{S}$, so that $\sigma'\sigma \notin \mathfrak{q}$. Since $\mathfrak{q}$ is prime, we have $a \in \mathfrak{q}$; that is, $\mathfrak{p} \subseteq \mathfrak{q}$. The reverse inclusion is proved similarly.

Let $\mathfrak{P}$ be a prime ideal in $S^{-1}R$. By part (i), there is some ideal I in R with $\mathfrak{P} = S^{-1}I$. We must show that I can be chosen to be a prime ideal in R. Now $h(R) \cap \mathfrak{P}$ is a prime ideal in $h(R)$, and so $\mathfrak{p} = h^{-1}(h(R) \cap \mathfrak{P})$ is a prime ideal in R. By part (i), $\mathfrak{P} = S^{-1}\mathfrak{p}$.

(v) If J is an ideal in $S^{-1}R$, then part (i) shows that $J = S^{-1}I$ for some ideal I in R. Since R is noetherian, we have $I = (r_1, \ldots, r_n)$, and so $J = (r_1/1, \ldots, r_n/1)$. Hence, every ideal in $S^{-1}R$ is finitely generated, and so $S^{-1}R$ is noetherian. •

Definition. If $\mathfrak{p}$ is a prime ideal in a commutative ring R, then the complement $S = R - \mathfrak{p}$ is multiplicatively closed, and $S^{-1}R$ is denoted by $R_\mathfrak{p}$.

Example 11.19.
If p is a prime in $\mathbb{Z}$, then $\mathfrak{p} = (p)$ is a prime ideal, and $\mathbb{Z}_\mathfrak{p} = \mathbb{Z}_{(p)}$. ◀

Proposition 11.20. *If R is a domain, then $\bigcap_\mathfrak{m} R_\mathfrak{m} = R$, where the intersection is over all the maximal ideals $\mathfrak{m}$ in R.*

Proof. Since R is a domain, $R_\mathfrak{m} \subseteq \text{Frac}(R)$ for all $\mathfrak{m}$, and so the intersection in the statement is defined. Moreover, it is plain that $R \subseteq R_\mathfrak{m}$ for all $\mathfrak{m}$, so that $R \subseteq \bigcap_m R_\mathfrak{m}$. For the reverse inclusion, let $a \in \bigcap R_\mathfrak{m}$. Define

$$I = (R : a) = \{r \in R : ra \in R\}.$$

If $I = R$, then $1 \in I$, and $a = 1a \in R$, as desired. If I is a proper ideal, then there exists a maximal ideal $\mathfrak{m}$ with $I \subseteq \mathfrak{m}$. Now $a/1 \in R_\mathfrak{m}$, so there is $r \in R$ and $\sigma \notin \mathfrak{m}$ with $a/1 = r/\sigma$; that is, $\sigma a = r \in R$. Hence, $\sigma \in I \subseteq \mathfrak{m}$, contradicting $\sigma \notin \mathfrak{m}$. Therefore, $R = \bigcap_\mathfrak{m} R_\mathfrak{m}$. •

The next proposition explains why $S^{-1}R$ is called localization.

Proposition 11.21. *If $\mathfrak{p}$ is a prime ideal in a commutative ring R, then $R_\mathfrak{p}$ is a local ring with maximal ideal $\mathfrak{p}R_\mathfrak{p} = \{r/s : r \in \mathfrak{p}$ and $s \notin \mathfrak{p}\}$.*

Proof. If $x \in R_\mathfrak{p}$, then $x = r/s$, where $r \in R$ and $s \notin \mathfrak{p}$. If $r \notin \mathfrak{p}$, then r/s is a unit in $R_\mathfrak{p}$; that is, all nonunits lie in $\mathfrak{p}R_\mathfrak{p}$. Hence, if I is any ideal in $R_\mathfrak{p}$ that contains an element r/s with $r \notin \mathfrak{p}$, then $I = R_\mathfrak{p}$. It follows that every proper ideal in $R_\mathfrak{p}$ is contained in $\mathfrak{p}R_\mathfrak{p}$, and so $R_\mathfrak{p}$ is a local ring with unique maximal ideal $\mathfrak{p}R_\mathfrak{p}$. •

The fundamental assumption underlying the local/global strategy is that the local case is simpler than the global. The structure of projective modules over a general ring can be quite complicated, but the next proposition shows that projective modules over local rings are free.

Lemma 11.22. *Let R be a local ring with maximal ideal $\mathfrak{m}$. An element $r \in R$ is a unit if and only if $r \notin \mathfrak{m}$.*

Proof. It is clear that if r is a unit, then $r \notin \mathfrak{m}$, for $\mathfrak{m}$ is a proper ideal. Conversely, assume that r is not a unit. By Zorn's lemma, there is a maximal ideal containing the principal ideal (r). Since R is local, there is only one maximal ideal, namely, $\mathfrak{m}$, and so $r \in \mathfrak{m}$. •

Proposition 11.23. *If R is a local ring, then every finitely generated*[3] *projective R-module B is free.*

Proof. Let R be a local ring with maximal ideal $\mathfrak{m}$, and let $\{b_1, \ldots, b_n\}$ be a *minimal* set of generators of B; that is, B cannot be generated by fewer than n elements. Let F be the

[3]It is a theorem of Kaplansky that the finiteness hypothesis can be omitted: Every projective module over a local ring is free. He even proves freeness when R is a noncommutative local ring.

free R-module with basis $x_1, \ldots, x_n$, and define $\varphi\colon F \to B$ by $\varphi(x_i) = b_i$ for all i. Thus, there is an exact sequence

$$0 \to K \to F \xrightarrow{\varphi} B \to 0, \tag{1}$$

where $K = \ker\varphi$.

We claim that $K \subseteq \mathfrak{m}F$. If, on the contrary, $K \subsetneq \mathfrak{m}F$, there is an element $y = \sum_{i=1}^{n} r_i x_i \in K$ which is not in $\mathfrak{m}F$; that is, some coefficient, say, $r_1 \notin \mathfrak{m}$. Now r_1 is a unit, by Lemma 11.22. Now $y \in K = \ker\varphi$ gives $\sum r_i b_i = 0$. Hence, $b_1 = -r_1^{-1}\big(\sum_{i=2}^{n} r_i b_i\big)$, which implies that $B = \langle b_2, \ldots, b_n\rangle$, contradicting the minimality of the original generating set.

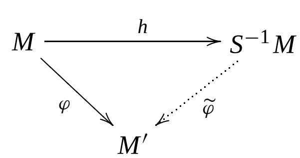
Returning to the exact sequence (1), projectivity of B gives $F = K \oplus B'$, where B' is a submodule of F with $B' \cong B$. Hence, $\mathfrak{m}F = \mathfrak{m}K \oplus \mathfrak{m}B'$. Since $\mathfrak{m}K \subseteq K \subseteq \mathfrak{m}F$, Corollary 7.18 gives

$$K = \mathfrak{m}K \oplus (K \cap \mathfrak{m}B').$$

But $K \cap \mathfrak{m}B' \subseteq K \cap B' = \{0\}$, so that $K = \mathfrak{m}K$. The submodule K is finitely generated, being a summand (and hence a homomorphic image) of the finitely generated module F, so that Nakayama's lemma (Corollary 8.32) gives $K = \{0\}$. Therefore, φ is an isomorphism and B is free. •

Having localized a commutative ring, we now localize its modules. If M is an R-module and $s \in R$, let μ_s denote the multiplication map $M \to M$ defined by $m \mapsto sm$. Note that if S is a subset of R, then $\mu_s\colon M \to M$ is invertible for every $s \in S$ if and only if M is an $S^{-1}R$-module.

Definition. Let R be a commutative ring and let S be any subset of R. A ***localization*** of an R-module M is an $S^{-1}R$-module $S^{-1}M$ (i.e.., $\mu_s\colon S^{-1}m \to S^{-1}M$ is invertible for all $s \in S$) and an R-map $h_M\colon M \to S^{-1}M$, called the ***localization map***, which is a solution to the following universal mapping problem:



If $\varphi\colon M \to M'$ is an R-map, where M' is an $S^{-1}R$-module, then there is a unique $S^{-1}R$-map $\widetilde{\varphi}\colon S^{-1}M \to M'$ making the diagram commute.

The obvious candidate for $S^{-1}M$—namely, $S^{-1}R \otimes_R M$—is, in fact, its localization.

Proposition 11.24. *Let R be a commutative ring, let S be any subset of R, and let M be an R-module. Then $S^{-1}R \otimes_R M$ and the R-map $h\colon M \to S^{-1}R \otimes_R M$, given by $m \mapsto 1 \otimes m$, is a localization of M.*

Proof. Let $\varphi\colon M \to M'$ be an R-map, where M' is an $S^{-1}R$-module. The function $S^{-1}R \times M \to M'$, defined by $(r/\sigma, m) \mapsto (r/\sigma)\varphi(m)$, where $r \in R$ and $\sigma \in \overline{S}$, is easily seen to be R-bilinear. Hence, there is a unique R-map $\widetilde{\varphi}\colon S^{-1}R \otimes_R M \to M'$ with $\widetilde{\varphi}h = \varphi$. Since $h(M)$ generates $S^{-1}R \otimes_R M$, $\widetilde{\varphi}$ is the unique R-map making the diagram commute. We let the reader check that $\widetilde{\varphi}$ is an $S^{-1}R$-map. •

One of the most important properties of $S^{-1}R$ is that it is flat as an R-module. To prove this, we first generalize the argument in Proposition 11.14.

Proposition 11.25. *If S is a subset of a commutative ring R, if M is an R-module, and if $h_M\colon M \to S^{-1}M$ is the localization map, then*

$$\ker h_M = \{m \in M : \sigma m = 0 \text{ for some } \sigma \in \overline{S}\}.$$

Proof. Denote $\{m \in M : \sigma m = 0 \text{ for some } \sigma \in \overline{S}\}$ by K. If $\sigma m = 0$, for $m \in M$ and $\sigma \in \overline{S}$, then $h_M(m) = (1/\sigma)h_M(\sigma m) = 0$, and so $K \subseteq \ker h_M$. For the reverse inclusion, proceed as in Proposition 11.14: If $m \in K$, there is $\sigma \in \overline{S}$ with $\sigma m = 0$. Reduce to the case $S = \{\sigma\}$ for some $\sigma \in \overline{S}$, so that $S^{-1}R = R[x]/(\sigma x - 1)$. Now $R[x] \otimes_R M \cong \sum_i Rx^i \otimes_R M$, because $R[x]$ is the free R-module with basis $\{1, x, x^2, \ldots\}$. Hence, each element in $R[x] \otimes_R M$ has a unique expression of the form $\sum_i x^i \otimes m_i$, where $m_i \in M$. In particular, if $m \in \ker h_M$, then

$$0 = 1 \otimes m = (\sigma x - 1)\sum_{i=0}^{n} x^i \otimes m_i = \sum_{i=0}^{n} (\sigma x^{i+1} \otimes m_i - x^i \otimes m_i).$$

The proof now finishes as the proof of Proposition 11.14. Expanding and equating coefficients gives equations

$$1 \otimes m = -1 \otimes m_0,\ x \otimes \sigma m_0 = x \otimes m_1,\ \ldots,$$
$$x^n \otimes \sigma m_{n-1} = x^n \otimes m_n,\ x^{n+1} \otimes \sigma m_n = 0.$$

It follows that

$$m = -m_0, \quad \sigma m_0 = m_1, \quad \ldots \quad \sigma m_{n-1} = m_n, \quad \sigma m_n = 0.$$

Hence, $\sigma m = -\sigma m_0 = -m_1$, and, by induction, $\sigma^i m = -m_i$ for all i. In particular, $\sigma^n m = -m_n$ and so $\sigma^{n+1} m = -\sigma m_n = 0$ in M. Therefore, $\ker h_M \subseteq K$, as desired. •

Corollary 11.26. *Let S be a subset of a commutative ring R and let M be an R-module.*

(i) *Every element $u \in S^{-1}M$ has the form $u = \sigma^{-1}m$ for some $\sigma \in \overline{S}$ and some $m \in M$.*

(ii) *$s_1^{-1}m_1 = s_2^{-1}m_2$ in $S^{-1}M$ if and only if $\sigma(s_1^{-1}m_1 - s_2^{-1}m_2)$ in M for some $\sigma \in \overline{S}$.*

Proof. (i) If $u \in S^{-1}M$, then $u = \sum_i (r_i/\sigma_i)m_i$, where $r_i \in R$, $\sigma_i \in \overline{S}$, and $m_i \in M$. If we define $\sigma = \prod \sigma_i$ and $\widehat{\sigma_i} = \prod_{j \neq i} \sigma_j$, then

$$\begin{aligned} u &= \sum (1/\sigma_i) r_i m_i \\ &= \sum (\widehat{\sigma_i}/\sigma) r_i m_i \\ &= (1/\sigma) \sum \widehat{\sigma_i} r_i m_i \\ &= (1/\sigma) m, \end{aligned}$$

where $m = \sum \widehat{\sigma_i} r_i m_i \in M$.

(ii) If $\sigma \in \overline{S}$ with $\sigma(s_2 m_1 - s_1 m_2) = 0$ in M, then $(\sigma/1)(s_2 m_1 - s_1 m_2) = 0$ in $S^{-1}M$. As $\sigma/1$ is a unit, $s_2 m_1 - s_1 m_2 = 0$, and so $s_1^{-1} m_1 = s_2^{-1} m_2$.

Conversely, if $s_1^{-1} m_1 = s_2^{-1} m_2$ in $S^{-1}M$, then $(1/s_1 s_2)(s_2 m_1 - s_1 m_2) = 0$. Since $1/s_1 s_2$ is a unit, we have $(s_2 m_1 - s_1 m_2) = 0$ and $s_2 m_1 - s_1 m_2 \in \ker h_M$. By Proposition 11.25, there exists $\sigma \in \overline{S}$ with $\sigma(s_2 m_1 - s_1 m_2) = 0$ in M. •

Corollary 11.27. *Let S be a subset of a commutative ring R. If A is an $S^{-1}R$-module, then $A \cong S^{-1}A$.*

Proof. Define $\varphi \colon A \to S^{-1}A$ by $a \mapsto 1 \otimes a$. If $\varphi(a) = 0$, then there is $\sigma \in \overline{S}$ with $\sigma a = 0$. Since σ is a unit in $S^{-1}R$ and A is an $S^{-1}R$-module, the equation $a = \sigma^{-1}\sigma a = 0$ makes sense in A. Hence, φ is an injection. To see that φ is a surjection, note that $(1/\sigma) \otimes a = \varphi(\sigma^{-1} a)$. •

Theorem 11.28. *If S is a subset of a commutative ring R, then $S^{-1}R$ is a flat R-module.*

Proof. We must show that if $0 \to A \xrightarrow{f} B$ is exact, then so is

$$0 \to S^{-1}R \otimes_R A \xrightarrow{1 \otimes f} S^{-1}R \otimes_R B.$$

Let $u \in \ker(1 \otimes f)$; by Corollary 11.26, $u = \sigma^{-1} \otimes a$ for some $\sigma \in \overline{S}$ and $a \in A$. Now $0 = (1 \otimes f)(u) = \sigma^{-1} \otimes f(a)$, so that $f(a) \in \ker h_M$. By Proposition 11.25, there is $\tau \in \overline{S}$ with $0 = \tau f a = f(\tau a)$. Thus, $\tau a \in \ker f = \{0\}$, because f is an injection. Therefore, $0 = 1 \otimes \tau a = \tau(1 \otimes a) = \tau u$. Finally, $u = 0$, because τ is a unit. Therefore, $1 \otimes f$ is an injection, and so $S^{-1}R$ is a flat R-module. •

Corollary 11.29. *If S is a subset of a commutative ring R, then localization $M \mapsto S^{-1}M = S^{-1}R \otimes_R M$ defines an exact functor ${}_R\mathbf{Mod} \to {}_{S^{-1}R}\mathbf{Mod}$.*

Proof. Localization is the functor $S^{-1}R \otimes_R$, and it is exact because $S^{-1}R$ is a flat R-module. •

Notation. In the special case $S = R - \mathfrak{p}$, where $\mathfrak{p}$ is a prime ideal in R, we write

$$S^{-1}M = M_{\mathfrak{p}}.$$

If $f : M \to N$ is an R-map, write $f_{\mathfrak{p}} : M_{\mathfrak{p}} \to N_{\mathfrak{p}}$, where $f_{\mathfrak{p}} = 1_{R_{\mathfrak{p}}} \otimes f$.

We restate Corollary 11.18(iv) in this notation. The function $\mathfrak{q} \mapsto \mathfrak{q}_{\mathfrak{p}}$ is a bijection from the family of all prime ideals in R that are contained in $\mathfrak{p}$ to $\operatorname{Spec}(R_{\mathfrak{p}})$ (see Exercise 6.67 on page 398).

Here are some globalization results.

Proposition 11.30. *Let I and J be ideals in a domain R. If $I_{\mathfrak{m}} = J_{\mathfrak{m}}$ for every maximal ideal $\mathfrak{m}$, then $I = J$.*

Proof. Take $b \in J$, and define

$$(I : b) = \{r \in R : rb \in I\}.$$

Let $\mathfrak{m}$ be a maximal ideal in R. Since $I_{\mathfrak{m}} = J_{\mathfrak{m}}$, there are $a \in I$ and $s \notin \mathfrak{m}$ with $b/1 = a/s$. As R is a domain, $sb = a \in I$, so that $s \in (I : b)$; but $s \notin \mathfrak{m}$, so that $(I : b) \subsetneq \mathfrak{m}$. Thus, $(I : b)$ cannot be a proper ideal, for it is not contained in any maximal ideal. Therefore, $(I : b) = R$; hence, $1 \in (I : b)$ and $b = 1b \in I$. We have proved that $J \subseteq I$, and the reverse inclusion is proved similarly. •

Proposition 11.31. *Let R be a commutative ring.*

(i) *If M is an R-module with $M_{\mathfrak{m}} = \{0\}$ for every maximal ideal $\mathfrak{m}$, then $M = \{0\}$.*

(ii) *If $f : M \to N$ is an R-map and $f_{\mathfrak{m}} : M_{\mathfrak{m}} \to N_{\mathfrak{m}}$ is an injection for every maximal ideal $\mathfrak{m}$, then f is an injection.*

(iii) *If $f : M \to N$ is an R-map and $f_{\mathfrak{m}} : M_{\mathfrak{m}} \to N_{\mathfrak{m}}$ is a surjection for every maximal ideal $\mathfrak{m}$, then f is a surjection.*

(iv) *If $f : M \to N$ is an R-map and $f_{\mathfrak{m}} : M_{\mathfrak{m}} \to N_{\mathfrak{m}}$ is an isomorphism for every maximal ideal $\mathfrak{m}$, then f is an isomorphism.*

Proof. (i) If $M \neq \{0\}$, then there is $m \in M$ with $m \neq 0$. It follows that the annihilator $I = \{r \in R : rm = 0\}$ is a proper ideal in R, for $1 \notin I$, and so there is some maximal ideal $\mathfrak{m}$ containing I. Now $1 \otimes m = 0$ in $M_{\mathfrak{m}}$, so that $m \in \ker h_M$. Proposition 11.25 gives $s \notin \mathfrak{m}$ with $sm = 0$ in M, for $R - \mathfrak{m}$ is multiplicatively closed. Hence, $s \in I \subseteq \mathfrak{m}$, and this is a contradiction. Therefore, $M = \{0\}$.

(ii) There is an exact sequence $0 \to K \to M \xrightarrow{f} N$, where $K = \ker f$. Since localization is an exact functor, there is an exact sequence

$$0 \to K_{\mathfrak{m}} \to M_{\mathfrak{m}} \xrightarrow{f_{\mathfrak{m}}} N_{\mathfrak{m}}$$

for every maximal ideal $\mathfrak{m}$. By hypothesis, each $f_{\mathfrak{m}}$ is an injection, so that $K_{\mathfrak{m}} = \{0\}$ for all maximal ideals $\mathfrak{m}$. Part (i) now shows that $K = \{0\}$, and so f is an injection.

(iii) There is an exact sequence $M \xrightarrow{f} N \to C \to 0$, where $C = \operatorname{coker} f = N/\operatorname{im} f$. Since tensor product is right exact, $C_{\mathfrak{m}} = \{0\}$ for all maximal ideals $\mathfrak{m}$, and so $C = \{0\}$. But f is surjective if and only if $C = \operatorname{coker} f = \{0\}$.

(iv) This follows at once from parts (ii) and (iii). •

We cannot weaken the hypothesis of Proposition 11.31(iv) to $M_{\mathfrak{m}} \cong N_{\mathfrak{m}}$ for all maximal ideals $\mathfrak{m}$; we must assume that all the local isomorphisms arise from a given map $f: M \to N$. If G is the subgroup of $\mathbb{Q}$ consisting of all a/b with b squarefree, then we saw, in Example 11.5, that $G_{(p)} \cong \mathbb{Z}_{(p)}$ for all primes p, but $G \ncong \mathbb{Z}$.

Exercises 11.20 and 11.22 on page 921 show that localization preserves projectives and flats; that is, if A is a projective R-module, then $S^{-1}A$ is a projective $(S^{-1}R)$-module, and if B is a flat R-module, then $S^{-1}B$ is a flat $(S^{-1}R)$-module. Preserving injectivity is more subtle.

Lemma 11.32. *Let S be a subset of a commutative ring R, and let M and A be R-modules with A finitely presented. Then there is a natural isomorphism*

$$\tau_A : S^{-1}\operatorname{Hom}_R(A, M) \to \operatorname{Hom}_{S^{-1}R}(S^{-1}A, S^{-1}M).$$

Proof. It suffices to construct natural isomorphisms

$$\theta_A : \operatorname{Hom}_R(A, S^{-1}M) \to \operatorname{Hom}_{S^{-1}R}(S^{-1}A, S^{-1}M)$$

and

$$\varphi_A : S^{-1}\operatorname{Hom}_R(A, M) \to \operatorname{Hom}_R(A, S^{-1}M),$$

for then we can define $\tau_A = \theta_A \varphi_A$.

Assume first that $A = R^n$ is a finitely generated free R-module. If $a_1, \ldots, a_n$ is a basis of A, then $a_1/1, \ldots, a_n/1$ is a basis of $S^{-1}A = S^{-1}R \otimes_R R^n$. The map

$$\theta_{R^n} : \operatorname{Hom}_R(A, S^{-1}M) \to \operatorname{Hom}_{S^{-1}R}(S^{-1}A, S^{-1}M),$$

given by $f \mapsto \widetilde{f}$, where $\widetilde{f}(a_i/\sigma) = f(a_i)/\sigma$, is easily seen to be a well-defined R-isomorphism.

If, now, A is a finitely presented R-module, then there is an exact sequence

$$R^t \to R^n \to A \to 0. \tag{2}$$

Applying the contravariant functors $\operatorname{Hom}_R(\ , M')$ and $\operatorname{Hom}_{S^{-1}R}(\ , M')$, where $M' = S^{-1}M$ is first viewed as an R-module, gives a commutative diagram with exact rows

$$\begin{array}{ccccccc}
0 \to & \operatorname{Hom}_R(A, M') & \longrightarrow & \operatorname{Hom}_R(R^n, M') & \longrightarrow & \operatorname{Hom}_R(R^t, M') & \\
& \downarrow{\scriptstyle \theta_A} & & \downarrow{\scriptstyle \theta_{R^n}} & & \downarrow{\scriptstyle \theta_{R^t}} & \\
0 \to & \operatorname{Hom}_{S^{-1}R}(S^{-1}A, M') & \to & \operatorname{Hom}_{S^{-1}R}((S^{-1}R)^n, M') & \to & \operatorname{Hom}_{S^{-1}R}((S^{-1}R)^t, M'). &
\end{array}$$

Since the vertical maps θ_{R^n} and θ_{R^t} are isomorphisms, there is a dotted arrow θ_A which must be an isomorphism, by Proposition 8.94. If $\beta \in \operatorname{Hom}_R(A, M)$, then the reader may check that

$$\theta_A(\beta) = \widetilde{\beta} : a/\sigma \mapsto \beta(a)/\sigma,$$

from which it follows that the isomorphisms θ_A are natural.

Construct $\varphi_A \colon S^{-1}\operatorname{Hom}_R(A, M) \to \operatorname{Hom}_R(A, S^{-1}M)$ by defining $\varphi_A \colon g/\sigma \mapsto g_\sigma$, where $g_\sigma(a) = g(a)/\sigma$. Note that φ_A is well-defined, for it arises from the R-bilinear function $S^{-1}R \times \operatorname{Hom}_R(A, M) \to \operatorname{Hom}_R(A, S^{-1}M)$ given by $(r/\sigma, g) \mapsto rg_\sigma$ (remember that $S^{-1}\operatorname{Hom}_R(A, M) = S^{-1}R \otimes_R \operatorname{Hom}_R(A, M)$). Observe that φ_A is an isomorphism when A is finitely generated free, and consider the commutative diagram

$$\begin{array}{ccccccc}
0 \longrightarrow & S^{-1}\operatorname{Hom}_R(A, M) & \longrightarrow & S^{-1}\operatorname{Hom}_R(R^n, M) & \longrightarrow & S^{-1}\operatorname{Hom}_R(R^t, M) \\
& \downarrow{\scriptstyle \varphi_A} & & \downarrow{\scriptstyle \varphi_{R^n}} & & \downarrow{\scriptstyle \varphi_{R^t}} \\
0 \longrightarrow & \operatorname{Hom}_R(A, S^{-1}M) & \longrightarrow & \operatorname{Hom}_R(R^n, S^{-1}M) & \longrightarrow & \operatorname{Hom}_R(R^t, S^{-1}M).
\end{array}$$

The top row is exact, for it arises from Eq. (2) by first applying the left exact contravariant functor $\operatorname{Hom}_R(\ , M)$, and then applying the exact localization functor. The bottom row is exact, for it arises from Eq. (2) by applying the left exact contravariant functor $\operatorname{Hom}_R(\ , S^{-1}M)$. The five lemma, Exercise 8.52 on page 604, shows that φ_A is an isomorphism. •

Example 11.33.
Lemma 11.32 can be false if A is not finitely presented. For example, let $R = \mathbb{Z}$ and $S^{-1}R = \mathbb{Q}$. We claim that

$$\mathbb{Q} \otimes_{\mathbb{Z}} \operatorname{Hom}_{\mathbb{Z}}(\mathbb{Q}, \mathbb{Z}) \ncong \operatorname{Hom}_{\mathbb{Q}}(\mathbb{Q} \otimes_{\mathbb{Z}} \mathbb{Q}, \mathbb{Q} \otimes_{\mathbb{Z}} \mathbb{Z}).$$

The left-hand side is $\{0\}$ because $\operatorname{Hom}_{\mathbb{Z}}(\mathbb{Q}, \mathbb{Z}) = \{0\}$. On the other hand, the right-hand side is $\operatorname{Hom}_{\mathbb{Z}}(\mathbb{Q}, \mathbb{Q}) \cong \mathbb{Q}$. ◂

Proposition 11.34. *If S is a subset of a commutative noetherian ring R, and if E is an injective R-module, then $S^{-1}E$ is an injective $(S^{-1}R)$-module.*

Remark. This result can fail if R is not noetherian. If k is a field and $R = k[X]$, where X is an uncountable set of variables, then there exists an injective R-module E and a subset S of R such that $S^{-1}E$ is not an injective $(S^{-1}R)$-module (see E. C. Dade, "Localization of Injective Modules," *Journal of Algebra* 69 (1981), 416–425). ◂

Proof. By the Baer criterion, Theorem 7.68, it suffices to prove that

$$i^* \colon \operatorname{Hom}_{S^{-1}R}(S^{-1}R, S^{-1}E) \to \operatorname{Hom}_{S^{-1}R}(J, S^{-1}E)$$

is surjective for every ideal J in $S^{-1}R$, where $i \colon J \to S^{-1}R$ is the inclusion. Now every ideal J in $S^{-1}R$ has the form $J = S^{-1}I$, by Corollary 11.18, where I is an ideal in R. Since R is noetherian, every ideal is a finitely presented R-module, and so Lemma 11.32

applies to give a commutative diagram whose vertical arrows are isomorphisms

$$\begin{array}{ccc} S^{-1}\operatorname{Hom}_R(R, E) & \longrightarrow & S^{-1}\operatorname{Hom}_R(I, E) \\ \downarrow{\scriptstyle \tau_R} & & \downarrow{\scriptstyle \tau_I} \\ \operatorname{Hom}_{S^{-1}R}(S^{-1}R, S^{-1}E) & \xrightarrow{i^*} & \operatorname{Hom}_{S^{-1}R}(S^{-1}I, S^{-1}E). \end{array}$$

Injectivity of E implies that $\operatorname{Hom}_R(R, E) \to \operatorname{Hom}_R(I, E)$ is surjective (where the arrow is induced from the inclusion $I \to R$), so that exactness of localization shows that the arrow in the top row is surjective. Since the vertical arrows are isomorphisms, the arrow in the bottom row is surjective. Therefore, $J = S^{-1}I$ is an injective $(S^{-1}R)$-module. •

Localization commutes with Tor, essentially because $S^{-1}R$ is a flat R-module.

Proposition 11.35. *If S is a subset of a commutative ring R, then there are isomorphisms*

$$S^{-1}\operatorname{Tor}_n^R(A, B) \cong \operatorname{Tor}_n^{S^{-1}R}(S^{-1}A, S^{-1}B)$$

for all $n \geq 0$ and for all R-modules A and B.

Proof. First consider the case $n = 0$. For fixed R-module A, there is a natural isomorphism

$$\tau_B \colon S^{-1}(A \otimes_R B) \to S^{-1}A \otimes_{S^{-1}R} S^{-1}B,$$

for either is a solution U of the universal mapping problem

$$\begin{array}{ccc} S^{-1}A \times S^{-1}B & \longrightarrow & U, \\ & {\scriptstyle f}\searrow \quad \swarrow{\scriptstyle \widetilde{f}} & \\ & M & \end{array}$$

where M is an $(S^{-1}R)$-module, f is $(S^{-1}R)$-bilinear, and $\widetilde{f}$ is an $(S^{-1}R)$-map.

If $\mathbf{P}_B$ is a deleted projective resolution of B, then exactness of localization, together with localization preserving projectives, show that $S^{-1}(\mathbf{P}_B)$ is a deleted projective resolution of $S^{-1}B$. Naturality of the isomorphisms τ_A gives an isomorphism of complexes

$$S^{-1}(A \otimes_R \mathbf{P}_B) \cong S^{-1}A \otimes_{S^{-1}R} S^{-1}(\mathbf{P}_B),$$

so that their homology groups are isomorphic. Since localization is an exact functor, Proposition 10.38 applies, and

$$H_n(S^{-1}(A \otimes_R \mathbf{P}_B)) \cong S^{-1}H_n(A \otimes_R \mathbf{P}_B) \cong S^{-1}\operatorname{Tor}_n^R(A, B).$$

On the other hand, since $S^{-1}(\mathbf{P}_B)$ is a deleted projective resolution of $S^{-1}B$, the definition of Tor gives

$$H_n(S^{-1}A \otimes_{S^{-1}R} S^{-1}(\mathbf{P}_B)) \cong \operatorname{Tor}_n^{S^{-1}R}(S^{-1}A, S^{-1}B). \quad \bullet$$

Corollary 11.36. *Let A be an R-module over a commutative ring R. If $A_\mathfrak{m}$ is a flat $R_\mathfrak{m}$-module for every maximal ideal $\mathfrak{m}$, then A is a flat R-module.*

Proof. The hypothesis, together with Proposition 10.96, give $\operatorname{Tor}_n^{R_\mathfrak{m}}(A_\mathfrak{m}, B_\mathfrak{m}) = \{0\}$ for all $n \geq 1$, for every R-module B, and for every maximal ideal $\mathfrak{m}$. But Proposition 11.35 gives $\operatorname{Tor}_n^R(A, B)_\mathfrak{m} = \{0\}$ for all maximal ideals $\mathfrak{m}$ and all $n \geq 1$. Finally, Proposition 11.31 shows that $\operatorname{Tor}_n^R(A, B) = \{0\}$ for all $n \geq 1$. Since this is true for all R-modules B, we have A flat. •

We must add some hypotheses to get a similar result for Ext (see Exercise 11.23 on page 921).

Lemma 11.37. *If R is a left noetherian ring and A is a finitely generated left R-module, then there is a projective resolution $\mathbf{P}_\bullet$ of A in which each P_n is finitely generated.*

Proof. Since A is finitely generated, there exists a finitely generated free left R-module P_0 and a surjective R-map $\varepsilon\colon P_0 \to A$. Since R is left noetherian, $\ker\varepsilon$ is finitely generated, and so there exists a finitely generated free left R-module P_1 and a surjective R-map $d_1\colon P_1 \to \ker\varepsilon$. If we define $D_1\colon P_1 \to P_0$ as the composite id_1, where $i\colon \ker\varepsilon \to P_0$ is the inclusion, then there is an exact sequence

$$0 \to \ker D_1 \to P_1 \xrightarrow{D_1} P_0 \xrightarrow{\varepsilon} A \to 0.$$

This construction can be iterated, for $\ker D_1$ is finitely generated, and the proof can be completed by induction. (We remark that we have, in fact, constructed a free resolution of A.) •

Proposition 11.38. *Let S be a subset of a commutative noetherian ring R. If A is a finitely generated R-module, then there are isomorphisms*

$$S^{-1}\operatorname{Ext}_R^n(A, B) \cong \operatorname{Ext}_{S^{-1}R}^n(S^{-1}A, S^{-1}B)$$

for all $n \geq 0$ and for all R-modules B.

Proof. Since R is noetherian and A is finitely generated, Lemma 11.37 says there is a projective resolution $\mathbf{P}$ of A each of whose terms is finitely generated. By Lemma 11.32, there is a natural isomorphism

$$\tau_A\colon S^{-1}\operatorname{Hom}_R(A, B) \to \operatorname{Hom}_{S^{-1}R}(S^{-1}A, S^{-1}B)$$

for every R-module B (a finitely generated module over a noetherian ring must be finitely presented). Now τ_A gives an isomorphism of complexes

$$S^{-1}(\operatorname{Hom}_R(\mathbf{P}_A, B)) \cong \operatorname{Hom}_{S^{-1}R}(S^{-1}(\mathbf{P}_A), S^{-1}B).$$

Taking homology of the left hand side gives

$$H_n(S^{-1}(\operatorname{Hom}_R(\mathbf{P}_A, B))) \cong S^{-1}H_n(\operatorname{Hom}_R(\mathbf{P}_A, B)) \cong S^{-1}\operatorname{Ext}_R^n(A, B),$$

because localization is an exact functor (Proposition 10.38). On the other hand, homology of the right hand side is

$$H_n(\mathrm{Hom}_{S^{-1}R}(S^{-1}(\mathbf{P}_A), S^{-1}B)) = \mathrm{Ext}^n_{S^{-1}R}(S^{-1}A, S^{-1}B),$$

because $S^{-1}(\mathbf{P}_A)$ is an $(S^{-1}R)$-projective resolution of $S^{-1}A$. •

Remark. An alternative proof of the Proposition 11.38 can be given using a deleted injective resolution $\mathbf{E}_B$ in the second variable. We must still assume that A is finitely generated, in order to use Lemma 11.32, but now we use the fact, when R is noetherian, that localization preserves injectives. ◀

Corollary 11.39. *Let A be a finitely generated R-module over a commutative noetherian ring R. Then $A_\mathfrak{m}$ is a projective $R_\mathfrak{m}$-module for every maximal ideal $\mathfrak{m}$ if and only if A is a projective R-module.*

Proof. Sufficiency is Exercise 11.20 on page 921, and necessity follows from Proposition 11.38: For every R-module B and maximal ideal $\mathfrak{m}$, we have

$$\mathrm{Ext}^1_R(A, B)_\mathfrak{m} \cong \mathrm{Ext}^1_{R_\mathfrak{m}}(A_\mathfrak{m}, B_\mathfrak{m}) = \{0\},$$

because $A_\mathfrak{m}$ is projective. By Proposition 11.31, $\mathrm{Ext}^1_R(A, B) = \{0\}$, which says that A is projective. •

EXERCISES

11.1 Prove that $\mathbb{Z}_{(p)} \not\cong \mathbb{Q}$ as $\mathbb{Z}_{(p)}$-modules.

11.2 If R is a domain with $Q = \mathrm{Frac}(R)$, prove that every R-subalgebra A of Q is a localization of R.
Hint. Define $S = \{b \in R : 1/b \in A\}$.

11.3 Prove that the following statements are equivalent for a torsion-free abelian group G of rank 1.

(i) G is finitely generated.
(ii) G is cyclic.
(iii) If $x \in G$ is nonzero, then $h_p(x) = 0$ for almost all p and $h_p(x) \neq \infty$ for all primes p.
(iv) $\tau(G) = \tau(\mathbb{Z})$.

11.4 (i) If G is a torsion-free abelian group of rank 1, prove that the additive group of $\mathrm{End}(G)$ is torsion-free of rank 1.
(ii) Let $x \in G$ be nonzero with $\chi(x) = (h_2(x), h_3(x), \ldots, h_p(x), \ldots)$, and let R be the subring of $\mathbb{Q}$ in which $\chi(1) = (k_2, k_3, \ldots, k_p, \ldots)$ and

$$k_p = \begin{cases} \infty & \text{if } h_p(x) = \infty \\ 0 & \text{if } h_p(x) \text{ is finite.} \end{cases}$$

Prove that $\mathrm{End}(G) \cong R$. Prove that there are infinitely many G with $\mathrm{Aut}(G) \cong \mathbb{I}_2$.

11.5 Let G and H be torsion-free abelian groups of rank 1.

(i) Prove that $G \otimes_{\mathbb{Z}} H$ is torsion-free of rank 1.

(ii) If (h_p) is the height sequence of a nonzero element $x \in G$, and if (k_p) is the height sequence of a nonzero element $y \in H$, prove that the height sequence of $x \otimes y$ is (m_p), where $m_p = h_p + k_p$ (we agree that $\infty + k_p = \infty$).

11.6 Let T be the set of all types, and define $\tau \leq \tau'$, for $\tau, \tau' \in T$, if there are height sequences $(k_p) \in \tau$ and $(k'_p) \in \tau'$ with $k_p \leq k'_p$ for all primes p.

(i) Prove that $\leq$ is a partial order on T.

(ii) Prove that if G and G' are torsion-free abelian groups of rank 1, then $\tau(G) \leq \tau(G')$ if and only if G is isomorphic to a subgroup of G'.

(iii) Prove that T is a lattice, and show that if $\tau = \tau(G)$ and $\tau' = \tau(G')$, then $\tau \wedge \tau' = \tau(G \cap G')$ and $\tau \vee \tau' = \tau(G + G')$.

(iv) If G and G' are torsion-free abelian groups of rank 1, prove that $\operatorname{Hom}(G, G') \neq \{0\}$ if and only if $\tau(G) \leq \tau(G')$.

11.7 If G is a p-primary abelian group, prove that G is a $\mathbb{Z}_{(p)}$-module.

11.8 If S is a subset of a commutative ring R, and if $\overline{S}$ is the multiplicatively closed subset it generates, prove that $(\overline{S})^{-1}R \cong S^{-1}R$.

11.9 If $S = \{s_1, \ldots, s_n\}$ is a finite nonempty subset of a commutative ring R, prove that $S^{-1}R \cong \{s\}^{-1}R$, where $s = s_1 \cdots s_n$.
Hint. If s^{-1} exists, then so does $s^{-1}(s_1 \cdots \widehat{s_i} \cdots s_n) = s_i^{-1}$.

11.10 Prove that every localization of a PID is a PID. Conclude that if $\mathfrak{p}$ is a prime ideal in a PID R, then $R_{\mathfrak{p}}$ is a DVR.

11.11 If R is a Boolean ring and $\mathfrak{m}$ is a maximal ideal in R, prove that $R_{\mathfrak{m}}$ is a field.

11.12 Let S be a subset of a commutative ring R, and let I and J be ideals in R.

(i) Prove that $S^{-1}(IJ) = (S^{-1}I)(S^{-1}J)$.

(ii) Prove that $S^{-1}(I : J) = (S^{-1}I : S^{-1}J)$.

11.13 A domain R is a ***valuation ring*** if, for all $a, b \in R$, either $a \mid b$ or $b \mid a$.

(i) Prove that every DVR is a valuation ring.

(ii) Let R be a domain with $F = \operatorname{Frac}(R)$. Prove that R is a valuation ring if and only if $a \in R$ or $a^{-1} \in R$ for each nonzero $a \in F$.

11.14 (i) Prove that every finitely generated ideal in a valuation ring is principal.

(ii) Prove that every finitely generated ideal in a valuation ring is projective.

11.15 An abelian group Γ is ***ordered*** if it is a partially ordered set in which $a+b \leq a'+b'$ whenever $a \leq a'$ and $b \leq b'$; call Γ a ***totally ordered abelian group*** if the partial order is a chain. A ***valuation*** on a field k is a function $v\colon k^\times \to \Gamma$, where Γ is a totally ordered abelian group, such that

$$v(ab) = v(a) + v(b);$$
$$v(a + b) \geq \min\{v(a), v(b)\}.$$

(i) If $a/b \in \mathbb{Q}$ is nonzero, write $a = p^m a'$ and $b = p^n b'$, where $m, n \geq 0$ and $(a', p) = 1 = (b', p)$. Prove that $v\colon \mathbb{Q}^\times \to \mathbb{Z}$, defined by $v(a/b) = m - n$, is a valuation.

(ii) If $v\colon k^\times \to \Gamma$ is a valuation on a field k, define $R = \{0\} \cup \{a \in k^\times : v(a) \geq 0\}$. Prove that R is a valuation ring. (Every valuation ring arises in this way from a suitable valuation on its fraction field. Moreover, the valuation ring is discrete when the totally ordered abelian group Γ is isomorphic to $\mathbb{Z}$.)

(iii) Prove that $a \in R$ is a unit if and only if $v(a) = 0$.

(iv) Prove that every valuation ring is a (not necessarily noetherian) local ring.

Hint. Show that $\mathfrak{m} = \{a \in R : v(a) > 0\}$ is the unique maximal ideal in R.

11.16 Let Γ be a totally ordered abelian group and let k be a field. Define $k[\Gamma]$ to be the group algebra (consisting of all functions $f : \Gamma \to k$ almost all of whose values are 0). As usual, if $f(\gamma) = r_\gamma$, we denote f by $\sum_{\gamma\in\Gamma} r_\gamma \gamma$.

(i) Define the ***degree*** of $f = \sum_{\gamma\in\Gamma} r_\gamma \gamma$ to be α if α is the largest index γ with $r_\gamma \neq 0$. Prove that $k[\Gamma]$ is a valuation ring, where $v(f)$ is the degree of f.

(ii) Give an example of a non-noetherian valuation ring.

11.17 A subset S of a commutative ring R is ***saturated*** if it is multiplicatively closed and $ab \in S$ implies $a \in S$ and $b \in S$.

(i) Prove that $U(R)$, the set of all units in R, is a saturated subset of R.

(ii) An element $r \in R$ is a ***zero divisor*** on an R-module A if there is some nonzero $a \in A$ with $ra = 0$. Prove that $Z(A)$, the set of all zero divisors on an R-module A, is a saturated subset of R.

(iii) If S is a multiplicatively closed subset of a commutative ring R, prove that there exists a unique smallest saturated subset S' containing S (we call S' the ***saturation*** of S). Prove that $(S')^{-1}R \cong S^{-1}R$.

(iv) Prove that a multiplicatively closed subset S is saturated if and only if its complement $R - S$ is a union of prime ideals.

11.18 Let S be a subset of a commutative ring R, and let M be a finitely generated R-module. Prove that $S^{-1}M = \{0\}$ if and only if there is $\sigma \in \overline{S}$ with $\sigma M = \{0\}$.

11.19 Let S be a subset of a commutative ring R, and let A be an R-module.

(i) If A is free, prove that $S^{-1}A$ is a free $(S^{-1}R)$-module.

(ii) If A is finitely generated, prove that $S^{-1}A$ is a finitely generated $(S^{-1}R)$-module.

(iii) If A is finitely presented, prove that $S^{-1}A$ is a finitely presented $(S^{-1}R)$-module.

11.20 If A is projective, prove that $S^{-1}A$ is a projective $(S^{-1}R)$-module.

11.21 If $\mathfrak{p}$ is a prime ideal in a commutative ring R and if A is a projective R-module, prove that $A_\mathfrak{p}$ is a free $R_\mathfrak{p}$-module.

11.22 If B is a flat R-module, where R is a commutative ring, prove that the localization $S^{-1}B$ is a flat $(S^{-1}R)$-module.

Hint. The composite of exact functors is exact.

11.23 (i) Give an example of an abelian group B for which $\text{Ext}^1_\mathbb{Z}(\mathbb{Q}, B) \neq \{0\}$.

(ii) Prove that $\mathbb{Q} \otimes_\mathbb{Z} \text{Ext}^1_\mathbb{Z}(\mathbb{Q}, B) \neq \{0\}$ for the abelian group B in part (i).

(iii) Prove that Proposition 11.38 may be false if R is noetherian but A is not finitely generated.

11.24 Let R be a commutative k-algebra, where k is a commutative ring, and let M be a k-module. Prove, for all $n \geq 0$, that

$$R \otimes_k \bigwedge^n(M) \cong \bigwedge^n(R \otimes_k M)$$

(of course, $\bigwedge^n(R \otimes_k M)$ means the nth exterior power of the R-module $R \otimes_k M$). Conclude, for all maximal ideals $\mathfrak{m}$ in k, that

$$\Big(\bigwedge^n(M)\Big)_{\mathfrak{m}} \cong \bigwedge^n(M_{\mathfrak{m}}).$$

Hint. Show that $R \otimes_k \bigwedge^n(M)$ is a solution to the universal mapping problem for alternating n-multilinear R-functions.

11.25 Let R be a commutative noetherian ring. If A and B are finitely generated R-modules, prove that $\mathrm{Tor}_n^R(A, B)$ and $\mathrm{Ext}_R^n(A, B)$ are finitely generated R-modules for all n.

11.2 DEDEKIND RINGS

A *Pythagorean triple* is a triple (a, b, c) of positive integers such that $a^2 + b^2 = c^2$. Examples of Pythagorean triples are $(3, 4, 5)$, $(5, 12, 13)$, and $(7, 24, 25)$, and all Pythagorean triples are classified in Exercise 1.23 on page 13 (there is an elegant geometric proof of this by Diophantus, ca. 100 AD). P. Fermat proved that there do not exist positive integers (a, b, c) with $a^4 + b^4 = c^4$ and, in 1637, he wrote in the margin of his copy of a book by Diophantus that he had a wonderful proof that there are no positive integers (a, b, c) with $a^n + b^n = c^n$ for any $n > 2$. Fermat's proof was never found, and his remark (that was merely a note to himself) became known only several years after Fermat's death, when Fermat's son published his father's works. There were other such statements left by Fermat, many of them true, some of them false, and this statement, the only one unresolved by 1800, was called Fermat's last theorem, perhaps in jest. It remained one of the outstanding challenges in number theory until 1995, when A. Wiles proved Fermat's last theorem.

Every positive integer $n > 2$ is a multiple of 4 or of some odd prime p. Thus, if there do not exist positive integers (a, b, c) with $a^p + b^p = c^p$ for every odd prime p, then Fermat's last theorem is true [if $n = pm$, then $a^n + b^n = c^n$ implies $(a^m)^p + (b^m)^p = (c^m)^p$]. Over the centuries, there were many attempts to prove it. For example, L. Euler published a proof (with gaps, later corrected) for the case $n = 3$, G. P. L. Dirichlet published a proof (with gaps, later corrected) for the case $n = 5$, and G. Lamé published a correct proof for the case $n = 7$.

The first major progress (not dealing only with particular primes p) was due to E. Kummer, in the middle of the 19th century. If $a^p + b^p = c^p$, where p is an odd prime, then a natural starting point of investigation is the identity

$$c^p = a^p + b^p = (a + b)(a + \zeta b)(a + \zeta^2 b) \cdots (a + \zeta^{p-1} b),$$

where $\zeta = \zeta_p$ is a primitive pth root of unity. Kummer proved that if $\mathbb{Z}[\zeta_p]$ is a UFD, where $\mathbb{Z}[\zeta_p] = \{f(\zeta_p) : f(x) \in \mathbb{Z}[x]\}$, then there do not exist positive integers a, b, c with $a^p + b^p = c^p$. On the other hand, he showed that there do exist primes p for which $\mathbb{Z}[\zeta_p]$ is not a UFD. To restore unique factorization, he invented "ideal numbers" that he

adjoined to $\mathbb{Z}[\zeta_p]$. Later, R. Dedekind recast Kummer's ideal numbers into our present notion of ideal. Thus, Fermat's last theorem has served as a catalyst in the development of both modern algebra and of algebraic number theory. *Dedekind rings* are the appropriate generalization of rings like $\mathbb{Z}[\zeta_p]$, and we will study them in this section.

Integrality

The notion of algebraic integer is a special case of the notion of integral element.

Definition. A ***ring extension*** R^*/R is a commutative ring R^* containing R as a subring. If R^*/R is a ring extension, then an element $a \in R^*$ is ***integral*** over R if it is a root of a monic polynomial in $R[x]$. A ring extension R^*/R is an ***integral extension*** if every $a \in R^*$ is integral over R.

Example 11.40.
The ***Noether Normalization Theorem*** is often used to prove the Nullstellensatz. It states that if k is a field and A is a finitely generated k-algebra, then there exist algebraically independent elements $a_1, \ldots, a_n$ in A so that A is integral over $k[a_1, \ldots, a_n]$. See Matsumura, *Commutative Ring Theory*, page 262. ◀

Recall that a complex number is an algebraic integer if it is a root of a monic polynomial in $\mathbb{Z}[x]$, so that algebraic integers are integral over $\mathbb{Z}$. The reader should compare the next lemma with Proposition 7.24.

Lemma 11.41. *If R^*/R is a ring extension, then the following conditions on a nonzero element $u \in R^*$ are equivalent.*

(i) *u is integral over R.*

(ii) *There is a finitely generated R-submodule B of R^* with $uB \subseteq B$.*

(iii) *There is a finitely generated faithful R-submodule B of R^* with $uB \subseteq B$; that is, if $dB = \{0\}$ for some $d \in R$, then $d = 0$.*

Proof. (i) ⇒ (ii). If u is integral over R, there is a monic polynomial $f(x) \in R[x]$ with $f(u) = 0$; that is, there are $r_i \in R$ with $u^n = \sum_{i=0}^{n-1} r_i u^i$. Define $B = \langle 1, u, u^2, \ldots, u^{n-1} \rangle$. It is clear that $uB \subseteq B$.

(ii) ⇒ (iii). If $B = \langle b_1, \ldots, b_m \rangle$ is a finitely generated R-submodule of R^* with $uB \subseteq B$, define $B' = \langle 1, b_1, \ldots, b_m \rangle$. Now B' is finitely generated, faithful (because $1 \in B'$), and $uB' \subseteq B'$.

(iii) ⇒ (i). Suppose there is a faithful R-submodule of R^*, say, $B = \langle b_i, \ldots, b_n \rangle$, with $uB \subseteq B$. There is a system of n equations $ub_i = \sum_{j=1}^{n} p_{ij} b_j$ with $p_{ij} \in R$. If $P = [p_{ij}]$ and if $X = (b_1, \ldots, b_n)^t$ is an $n \times 1$ column vector, then the $n \times n$ system can be rewritten in matrix notation: $(uI - P)X = 0$. Now $0 = \big(\operatorname{adj}(uI - P)\big)(uI - P)X = dX$, where $d = \det(uI - P)$, by Corollary 9.161. Since $dX = 0$, we have $db_i = 0$ for all i, and so

$dB = \{0\}$. Therefore, $d = 0$, because B is faithful. On the other hand, Corollary 9.154 gives $d = f(u)$, where $f(x) \in R[x]$ is a monic polynomial of degree n; hence, u is integral over R. •

Being an integral extension is transitive.

Proposition 11.42. *If $T \subseteq S \subseteq R$ are commutative rings with S integral over T and R integral over S, then R is integral over T.*

Proof. If $r \in R$, there is an equation $r^n + s_{n-1}r^{n-1} + \cdots + r_0 = 0$, where $s_i \in S$ for all i. By Lemma 11.41, the subring $S' = T[s_{n-1}, \ldots, s_0]$ is a finitely generated T-module. But r is integral over S', so that the ring $S'[r]$ is a finitely generated S'-module. Therefore, $S'[r]$ is a finitely generated T-module, and so r is integral over T. •

Proposition 11.43. *Let E/R be a ring extension.*

(i) *If $u, v \in E$ are integral over R, then both uv and $u + v$ are integral over R.*

(ii) *The commutative ring $\mathcal{O}_{E/R}$, defined by*

$$\mathcal{O}_{E/R} = \{u \in E : u \text{ is integral over } R\},$$

is an R-subalgebra of E.

Proof. (i) Since u and v are integral over R, Lemma 11.41(ii) says there are R-submodules $B = \langle b_1, \ldots, b_n \rangle$ and $C = \langle c_1, \ldots, c_m \rangle$ of E with $uB \subseteq B$ and $vC \subseteq C$; that is, $ub_i \in B$ for all i and $vc_j \in C$ for all j. Define BC to be the R-submodule of E generated by all $b_i c_j$; of course, BC is finitely generated. Now $uvBC \subseteq BC$, for $uvb_ic_j = (ub_i)(vc_j)$ is an R-linear combination of $b_k c_\ell$s, and so uv is integral over R. Similarly, $u + v$ is integral over R, for $(u + v)b_ic_j = (ub_i)c_j + (vc_j)b_i \in BC$.

(ii) Part (i) shows that $\mathcal{O}_{E/R}$ is closed under multiplication and addition. Now $R \subseteq \mathcal{O}_E$, for if $r \in R$, then r is a root of $x - r$. It follows that $1 \in \mathcal{O}_{E/R}$ and that $\mathcal{O}_{E/R}$ is an R-subalgebra of E. •

Here is a second proof of Proposition 11.43(i) for a domain E which uses tensor products and linear algebra. Let $f(x) \in R[x]$ be the minimal polynomial of u, let A be the companion matrix of $f(x)$, and let y be an eigenvector [over the algebraic closure of $\mathrm{Frac}(E)$]: $Ay = uy$. Let $g(x)$ be the minimal polynomial of v, let B be the companion matrix of $g(x)$, and let $Bz = vz$. Now

$$(A \otimes B)(y \otimes z) = Ay \otimes Bz = uy \otimes vz = uv(y \otimes z).$$

Therefore, uv is an eigenvalue of $A \otimes B$; that is, uv is a root of the monic polynomial $\det(xI - A \otimes B)$, which lies in $R[x]$ because both A and B have all their entries in R. Therefore, uv is integral over R. Similarly, the equation

$$(A \otimes I + I \otimes B)(y \otimes z) = Ay \otimes z + y \otimes Bz = (u + v)y \otimes z$$

shows that $u + v$ is integral over R. •

Definition. Let E/R be a ring extension. The R-subalgebra $\mathcal{O}_{E/R}$ of E, consisting of all those elements integral over R, is called the ***integral closure*** of R in E. If $\mathcal{O}_{E/R} = R$, then R is called ***integrally closed in*** E. If R is a domain and R is integrally closed in $F = \text{Frac}(R)$, that is, $\mathcal{O}_{F/R} = R$, then R is called ***integrally closed***.

Thus, R is integrally closed if $\alpha \in \text{Frac}(R)$ and α is integral over R, then $\alpha \in R$.

Example 11.44.
The ring $\mathcal{O}_{\mathbb{Q}/\mathbb{Z}} = \mathbb{Z}$, for if a rational number a is a root of a monic polynomial in $\mathbb{Z}[x]$, then Theorem 3.43 shows that $a \in \mathbb{Z}$. Hence, $\mathbb{Z}$ is integrally closed. ◄

Proposition 11.45. *Every* UFD R *is integrally closed. In particular, every* PID *is integrally closed.*

Proof. Let $F = \text{Frac}(R)$, and suppose that $u \in F$ is integral over R. Thus, there is an equation

$$u^n + r_{n-1}u^{n-1} + \cdots + r_1 u + r_0 = 0,$$

where $r_i \in R$. We may write $u = b/c$, where $b, c \in R$ and $(b, c) = 1$ (gcd's exist because R is a UFD, and so every fraction can be put in lowest terms). Substituting and clearing denominators,

$$b^n + r_{n-1}b^{n-1}c + \cdots + r_1 bc^{n-1} + r_0 c^n = 0.$$

Hence, $b^n = -c\big(r_{n-1}b^{n-1} + \cdots + r_1 bc^{n-2} + r_0 c^{n-1}\big)$, so that $c \mid b^n$ in R. But $(b, c) = 1$ implies $(b^n, c) = 1$, so that c must be a unit in R; that is, $c^{-1} \in R$. Therefore, $u = b/c = bc^{-1} \in R$, and so R is integrally closed. •

We now understand Example 6.21. If k is a field, the subring R of $k[x]$, consisting of all polynomials $f(x) \in k[x]$ having no linear term, is not a UFD because it is not integrally closed. It is easy to check that $\text{Frac}(R) = k(x)$, for $x = x^3/x^2 \in \text{Frac}(R)$. But $x \in k(x)$ is a root of the monic polynomial $t^2 - x^2 \in R[t]$, and $x \notin R$.

Definition. An ***algebraic number field*** is a finite field extension of $\mathbb{Q}$. If E is an algebraic number field, then $\mathcal{O}_{E/\mathbb{Z}}$ is usually denoted by $\mathcal{O}_E$ instead of by $\mathcal{O}_{E/\mathbb{Z}}$, and it is called the ***ring of integers*** in E.

Because of this new use of the word *integers*, algebraic number theorists often speak of the ring of ***rational integers*** when referring to $\mathbb{Z}$.

Proposition 11.46. *Let E be an algebraic number field and let $\mathcal{O}_E$ be its ring of integers.*

(i) *If $\alpha \in E$, there there is a nonzero integer m with $m\alpha \in \mathcal{O}_E$.*

(ii) $\text{Frac}(\mathcal{O}_E) = E$.

(iii) *$\mathcal{O}_E$ is integrally closed.*

Proof. (i) If $\alpha \in E$, then there is a monic polynomial $f(x) \in \mathbb{Q}[x]$ with $f(\alpha) = 0$. Clearing denominators gives an integer m with

$$m\alpha^n + c_{n-1}\alpha^{n-1} + c_{n-2}\alpha^{n-2} + \cdots + c_1\alpha + c_0 = 0,$$

where all $c_i \in \mathbb{Z}$. Multiplying by m^{n-1} gives

$$(m\alpha)^n + c_{n-1}(m\alpha)^{n-1} + mc_{n-2}(m\alpha)^{n-2} + \cdots + c_1 m^{n-2}(m\alpha) + m^{n-1}c_0 = 0.$$

Thus, $m\alpha \in \mathcal{O}_E$.

(ii) It suffices to show that if $\alpha \in E$, then there are $a, b \in \mathcal{O}_E$ with $\alpha = a/b$. But $m\alpha \in \mathcal{O}_E$, by part (i), $m \in \mathbb{Z} \subseteq \mathcal{O}_E$, and $\alpha = (m\alpha)/m$.

(iii) Suppose that $\alpha \in \text{Frac}(\mathcal{O}_E) = E$ is integral over $\mathcal{O}_E$. By transitivity of integral extensions, Proposition 11.42, we have α integral over $\mathbb{Z}$. But this means that $\alpha \in \mathcal{O}_E$ which is, by definition, the set of all those elements in E that are integral over $\mathbb{Z}$. Therefore, $\mathcal{O}_E$ is integrally closed. •

Example 11.47.
We shall see, in Proposition 11.76, that if $E = \mathbb{Q}(i)$, then $\mathcal{O}_E = \mathbb{Z}[i]$, the Gaussian integers. Now $\mathbb{Z}[i]$ is a PID, because it is a euclidean ring, and hence it is a UFD. The generalization of this example which replaces $\mathbb{Q}(i)$ by an algebraic number field E is more subtle. It is true that $\mathcal{O}_E$ is integrally closed, but it may not be not true that the elements of $\mathcal{O}_E$ are $\mathbb{Z}$-linear combinations of α. Moreover, the rings $\mathcal{O}_E$ may not be UFDs. We will investigate rings of integers at the end of this section. ◂

Given a ring extension R^*/R, what is the relation between ideals in R^* and ideals in R?

Definition. Let R^*/R be a ring extension. If I is an ideal in R, define its ***extension*** I^e to be R^*I, the ideal in R^* generated by I. If I^* is an ideal in R^*, define its ***contraction*** $I^{*c} = R \cap I^*$.

Remark. The definition can be generalized. Let $h\colon R \to R^*$ be a ring homomorphism, where R and R^* are any two commutative rings. Define the extension of an ideal I in R to be the ideal in R^* generated by $h(I)$; define the contraction of an ideal I^* in R^* to be $h^{-1}(I^*)$. If R^*/R is a ring extension, then taking $h\colon R \to R^*$ to be the inclusion gives the definition above. Another interesting instance is the localization map $h\colon R \to S^{-1}R$. ◂

Example 11.48.
(i) In general, the contraction function $c\colon \text{Spec}(R^*) \to \text{Spec}(R)$ is neither an injection nor a surjection. For example, $c\colon \text{Spec}(\mathbb{Q}) \to \text{Spec}(\mathbb{Z})$ is not surjective, while $c\colon \text{Spec}(\mathbb{Q}[x]) \to \text{Spec}(\mathbb{Q})$ is not injective.

(ii) It is easy to see that if R^*/R is a ring extension and $\mathfrak{p}^*$ is a prime ideal in R^*, then its contraction $\mathfrak{p}^* \cap R$ is also a prime ideal. If $a, b \in R$ and $ab \in \mathfrak{p}^* \cap R \subseteq \mathfrak{p}^*$, then $\mathfrak{p}^*$ prime

gives $a \in \mathfrak{p}^*$ or $b \in \mathfrak{p}^*$; as $a, b \in R$, either $a \in \mathfrak{p}^* \cap R$ or $b \in \mathfrak{p}^* \cap R$. Thus, contraction defines a function $c\colon \operatorname{Spec}(R^*) \to \operatorname{Spec}(R)$.

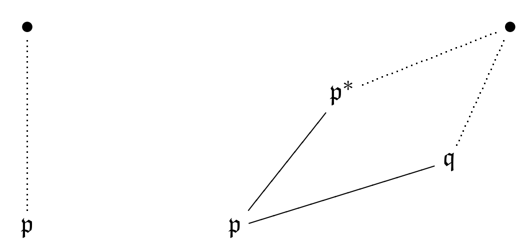
(iii) The contraction of a maximal ideal, though necessarily prime, need not be maximal. For example, if R^* is a field, then $\{0\}^*$ is a maximal ideal in R^*, but if R is not a field, then the contraction of $\{0\}^*$, namely, $\{0\}$, is not a maximal ideal in R. ◄

Example 11.49.
(i) Let $\mathcal{I}(R)$ denote the family of all the ideals in a commutative ring R. Extension defines a function $e\colon \mathcal{I}(R) \to \mathcal{I}(R^*)$; in general, it is neither injective nor surjective. If R^* is a field and R is not a field, then $e\colon \mathcal{I}(R) \to \mathcal{I}(R^*)$ is not injective; if R is a field and R^* is not a field, then $e\colon \mathcal{I}(R) \to \mathcal{I}(R^*)$ is not surjective.

(ii) If R^*/R is a ring extension and $\mathfrak{p}$ is a prime ideal in R, then its extension $R^*\mathfrak{p}$ need not be a prime ideal. Observe first that if $(a) = Ra$ is a principal ideal in R, then its extension is the principal ideal R^*a in R^* generated by a. Now let $R = \mathbb{R}[x]$ and $R^* = \mathbb{C}[x]$. The ideal $(x^2 + 1)$ is prime, because $x^2 + 1$ is irreducible in $\mathbb{R}[x]$, but its extension is not prime because $x^2 + 1$ factors in $\mathbb{C}[x]$. ◄

There are various elementary properties of extension and contraction, such as $I^{*ce} \subseteq I^*$ and $I^{ec} \supseteq I$, that are collected in Exercise 11.28 on page 930.

Is there a reasonable condition on a ring extension R^*/R that will give a good relationship between prime ideals in R and prime ideals in R^*? This question was posed and answered by I. S. Cohen and A. Seidenberg. We say that a ring extension R^*/R satisfies ***lying over*** if, for every prime ideal $\mathfrak{p}$ in R, there exists a prime ideal $\mathfrak{p}^*$ in R^* with $\mathfrak{p}^* \cap R = \mathfrak{p}$. We say that R^*/R satisfies ***going up*** if $\mathfrak{p} \subseteq \mathfrak{q}$ are prime ideals in R and if $\mathfrak{p}^*$ lies over $\mathfrak{p}$, then there exists a prime ideal $\mathfrak{q}^* \supseteq \mathfrak{p}^*$ which lies over $\mathfrak{q}$.



Lying over Going Up

We are going to see that extension and contraction are well-behaved in the presence of integral extensions.

Lemma 11.50. *Let R^* be an integral extension of R.*

(i) *If $\mathfrak{p}$ is a prime ideal in R and if $\mathfrak{p}^*$ lies over $\mathfrak{p}$, then $R^*/\mathfrak{p}^*$ is integral over $R/\mathfrak{p}$.*

(ii) *Is S is a subset of R, then $S^{-1}R^*$ is integral over $S^{-1}R$.*

Proof. (i) First, the second isomorphism theorem allows us to regard $R/\mathfrak{p}$ as a subring of $R^*/\mathfrak{p}^*$:

$$R/\mathfrak{p} = R/(\mathfrak{p}^* \cap R) \cong (R + \mathfrak{p}^*)/\mathfrak{p}^* \subseteq R^*/\mathfrak{p}^*.$$

Each element in $R^*/\mathfrak{p}^*$ has the form $\alpha + \mathfrak{p}^*$, where $\alpha \in R^*$. Since R^* is integral over R, there is an equation

$$\alpha^n + r_{n-1}\alpha^{n-1} + \cdots + r_0 = 0,$$

where $r_i \in R$. Now view this equation mod $\mathfrak{p}^*$ to see that $\alpha + \mathfrak{p}^*$ is integral over $R/\mathfrak{p}$.

(ii) If $\alpha^* \in S^{-1}R^*$, then $\alpha^* = \alpha/\sigma$, where $\alpha \in R^*$ and $\sigma \in \overline{S}$. Since R^* is integral over R, there is an equation $\alpha^n + r_{n-1}\alpha^{n-1} + \cdots + r_0 = 0$ with $r_i \in R$. Multiplying by $1/\sigma^n$ in $S^{-1}R^*$ gives

$$(\alpha/\sigma)^n + (r_{n-1}/\sigma)(\alpha/\sigma)^{n-1} + \cdots + r_0/\sigma^n = 0,$$

which shows that α/σ is integral over $S^{-1}R$. •

When R^*/R is a ring extension and R is a field, every proper ideal in R^* contracts to $\{0\}$ in R. The following proposition eliminates this collapse when R^* is an integral extension of R.

Proposition 11.51. *Let R^*/R be a ring extension of domains with R^* integral over R. Then R^* is a field if and only if R is a field.*

Proof. Assume that R^* is a field. If $u \in R$ is nonzero, then $u^{-1} \in R^*$, and so u^{-1} is integral over R. Therefore, there is an equation $(u^{-1})^n + r_{n-1}(u^{n-1})^{n-1} + \cdots + r_0 = 0$, where the $r_i \in R$. Multiplying by u^n gives $u^{-1} = -u(r_{n-1} + \cdots + r_0 u^{n-1})$. Therefore, $u^{-1} \in R$ and R is a field.

Conversely, assume that R is a field. If $\alpha \in R^*$ is nonzero, then there is a monic $f(x) \in R[x]$ with $f(\alpha) = 0$. Thus, α is algebraic over R, and so we may assume that $f(x) = \mathrm{irr}(\alpha, R)$; that is, $f(x)$ is irreducible. If $f(x) = \sum_{i=0}^{n} r_i x^i$, then

$$\alpha(\alpha^{n-1} + r_{n-1}\alpha^{n-1} + \cdots + r_1) = -r_0.$$

Irreducibility of $f(x)$ gives $r_0 \neq 0$, so that α^{-1} lies in R^*. Therefore, R^* is a field. •

Corollary 11.52. *Let R^*/R be an integral extension. If $\mathfrak{p}$ is a prime ideal in R and $\mathfrak{p}^*$ is a prime ideal lying over $\mathfrak{p}$, then $\mathfrak{p}$ is a maximal ideal if and only if $\mathfrak{p}^*$ is a maximal ideal.*

Proof. By Lemma 11.50(i), the domain $R^*/\mathfrak{p}^*$ is integral over the domain $R/\mathfrak{p}$. But now Proposition 11.51 says that $R^*/\mathfrak{p}^*$ is a field if and only if $R/\mathfrak{p}$ is a field; that is, $\mathfrak{p}^*$ is a maximal ideal in R^* if and only if $\mathfrak{p}$ is a maximal ideal in R. •

Corollary 11.53. *If E is an algebraic number field, then every nonzero prime ideal in $\mathcal{O}_E$ is a maximal ideal.*

Proof. Let $\mathfrak{p}$ be a nonzero prime ideal in $\mathcal{O}_E$. If $\mathfrak{p} \cap \mathbb{Z} \neq \{0\}$, then there is a prime p with $\mathfrak{p} \cap \mathbb{Z} = (p)$, by Example 11.48(i). But (p) is a maximal ideal in $\mathbb{Z}$, so that $\mathfrak{p}$ is a maximal ideal, by Corollary 11.52. It remains to show that $\mathfrak{p} \cap \mathbb{Z} \neq \{0\}$. Let $\alpha \in \mathfrak{p}$ be nonzero. Since α is integral over $\mathbb{Z}$, there is an equation

$$\alpha^n + c_{n-1}\alpha^{n-1} + \cdots + c_1\alpha + c_0 = 0,$$

where $c_i \in \mathbb{Z}$ for all i. If we choose such an equation with n minimal, then $c_0 \neq 0$. Since $\alpha \in \mathfrak{p}$, we have $c_0 = -\alpha(\alpha_{n-1} + c_{n-1}\alpha^{n-2} + \cdots + c_1) \in \mathfrak{p} \cap \mathbb{Z}$, so that $\mathfrak{p} \cap \mathbb{Z}$ is nonzero. •

Corollary 11.54. *Let R^* be integral over R, let $\mathfrak{p}$ be a prime ideal in R, and let $\mathfrak{p}^*$ and $\mathfrak{q}^*$ be prime ideals in R^* lying over $\mathfrak{p}$. If $\mathfrak{p}^* \subseteq \mathfrak{q}^*$, then $\mathfrak{p}^* = \mathfrak{q}^*$.*

Proof. Lemma 11.50(ii) and Corollary 11.18(iii) show that the hypotheses are preserved by localizing at $\mathfrak{p}$; that is, $R^*_{\mathfrak{p}}$ is integral over $R_{\mathfrak{p}}$ and $\mathfrak{p}^* R^*_{\mathfrak{p}} \subseteq \mathfrak{q}^* R^*_{\mathfrak{p}}$ are prime ideals. Hence, replacing R^* and R by their localizations, we may assume that R^* and R are local rings and that $\mathfrak{p}$ is a maximal ideal in R (by Proposition 11.21). But Corollary 11.52 says that maximality of $\mathfrak{p}$ forces maximality of $\mathfrak{p}^*$. Since $\mathfrak{p}^* \subseteq \mathfrak{q}^*$, we have $\mathfrak{p}^* = \mathfrak{q}^*$. •

Here are the theorems of Cohen and Seidenberg.

Theorem 11.55 (Lying Over). *Let R^*/R be a ring extension with R^* integral over R. If $\mathfrak{p}$ is a prime ideal in R, then there is a prime ideal $\mathfrak{p}^*$ in R^* lying over $\mathfrak{p}$; that is, $\mathfrak{p}^* \cap R = \mathfrak{p}$.*

Proof. There is a commutative diagram

$$\begin{array}{ccc} R & \xrightarrow{\ i\ } & R^*, \\ {\scriptstyle h}\downarrow & & \downarrow{\scriptstyle h^*} \\ R_{\mathfrak{p}} & \xrightarrow[\ j\ ]{} & S^{-1}R^* \end{array}$$

where h and h^* are localization maps and i and j are inclusions. If $S = R - \mathfrak{p}$, then $S^{-1}R^*$ is an extension of $R_{\mathfrak{p}}$ (since localization is an exact functor, R contained in R^* implies $R_{\mathfrak{p}}$ contained in $S^{-1}R^*$); by Lemma 11.50, $S^{-1}R^*$ is integral over $R_{\mathfrak{p}}$. Choose a maximal ideal $\mathfrak{m}^*$ in $S^{-1}R^*$. By Corollary 11.52, $\mathfrak{m}^* \cap R_{\mathfrak{p}}$ is a maximal ideal in $R_{\mathfrak{p}}$. But $R_{\mathfrak{p}}$ is a local ring with unique maximal ideal $\mathfrak{p}R_{\mathfrak{p}}$, so that $\mathfrak{m}^* \cap R_{\mathfrak{p}} = \mathfrak{p}R_{\mathfrak{p}}$. Since the inverse image of a prime ideal (under any ring map) is always prime, the ideal $\mathfrak{p}^* = (h^*)^{-1}(\mathfrak{m}^*)$ is a prime ideal in R^*. Now

$$(h^*i)^{-1}(\mathfrak{m}^*) = i^{-1}(h^*)^{-1}(\mathfrak{m}^*) = i^{-1}(\mathfrak{p}^*) = \mathfrak{p}^* \cap R,$$

while

$$(jh)^{-1}(\mathfrak{m}^*) = h^{-1}j^{-1}(\mathfrak{m}^*) = h^{-1}(\mathfrak{m}^* \cap R_{\mathfrak{p}}) = h^{-1}(\mathfrak{p}R_{\mathfrak{p}}) = \mathfrak{p}.$$

Therefore, $\mathfrak{p}^*$ is a prime ideal lying over $\mathfrak{p}$. •

Theorem 11.56 (Going Up). *Let R^*/R be a ring extension with R^* integral over R. If $\mathfrak{p} \subseteq \mathfrak{q}$ are prime ideals in R, and if $\mathfrak{p}^*$ is a prime ideal in R^* lying over $\mathfrak{p}$, then there exists a prime ideal $\mathfrak{q}^*$ lying over $\mathfrak{q}$ with $\mathfrak{p}^* \subseteq \mathfrak{q}^*$.*

Proof. Lemma 11.50 says that $(R^*/\mathfrak{p}^*)/(R/\mathfrak{p})$ is an integral ring extension, where $R/\mathfrak{p}$ is imbedded in $R^*/\mathfrak{p}^*$ as $(R + \mathfrak{p}^*)/\mathfrak{p}^*$. Replacing R^* and R by these quotient rings, we may assume that both $\mathfrak{p}^*$ and $\mathfrak{p}$ are $\{0\}$. The theorem now follows at once from the lying over theorem. •

There is also a going down theorem, but it requires an additional hypothesis.

Theorem (Going Down). *Let R^*/R be an integral extension and assume that R is integrally closed. If $\mathfrak{p}_1 \supseteq \mathfrak{p}_2 \supseteq \cdots \supseteq \mathfrak{p}_n$ is a chain of prime ideals in R and if $\mathfrak{p}_1^* \supseteq \mathfrak{p}_2^* \supseteq \cdots \supseteq \mathfrak{p}_m^*$, for $m < n$ is a chain of prime ideals in R^* with each $\mathfrak{p}_i^*$ lying over $\mathfrak{p}_i$, then the chain in R^* can be extended to $\mathfrak{p}_1^* \supseteq \mathfrak{p}_2^* \supseteq \cdots \supseteq \mathfrak{p}_n^*$ with $\mathfrak{p}_i^*$ lying over $\mathfrak{p}_i$ for all $i \leq n$.*

Proof. See Atiyah-Macdonald, *Introduction to Commutative Algebra*, page 64. •

EXERCISES

11.26 If R is an integrally closed domain and S is a multiplicatively closed subset of R not containing 0, prove that $S^{-1}R$ is also integrally closed.

11.27 Prove that every valuation ring is integrally closed.

11.28 Let R^*/R be a ring extension. If I is an ideal in R, denote its extension by I^e; if I^* is an ideal in R^*, denote its contraction by I^{*c}. Prove each of the follow assertions.

(i) Both e and c preserve inclusion: If $I \subseteq J$, then $I^e \subseteq J^e$; if $I^* \subseteq J^*$, then $I^{*c} \subseteq J^{*c}$.

(ii) $I^{*ce} \subseteq I^*$ and $I^{ec} \supseteq I$.

(iii) $I^{*cec} = I^{*c}$ and $I^{ece} = I^e$.

(iv) $(I^* + J^*)^c \supseteq I^{*c} + J^{*c}$ and $(I + J)^e = I^e + J^e$.

(v) $(I^* \cap J^*)^c = I^{*c} \cap J^{*c}$ and $(I \cap J)^e \subseteq I^e \cap J^e$.

(vi) $(I^*J^*)^c \supseteq I^{*c}J^{*c}$ and $(IJ)^e = I^eJ^e$.

(vii) $\big(\sqrt{I^*}\big)^c = \sqrt{I^{*c}}$ and $\big(\sqrt{I}\big)^e \subseteq \sqrt{I^e}$.

(viii) $(J^* : I^*)^c \subseteq (J^{*c} : I^{*c})$ and $(I : J)^e \subseteq (I^e : J^e)$.

11.29 If $\mathbb{A}$ is the field of all algebraic numbers, then $\mathcal{O}_{\mathbb{A}}$ is the ring of all algebraic integers. Prove that
$$\mathcal{O}_{\mathbb{A}} \cap \mathbb{Q} = \mathbb{Z}.$$
Conclude, for every algebraic number field E, that $\mathcal{O}_E \cap \mathbb{Q} = \mathbb{Z}$.

11.30 Let R^*/R be an integral ring extension.

(i) If $a \in R$ is a unit in R^*, prove that a is a unit in R.

(ii) Prove that $J(R) = R \cap J(R^*)$, where $J(R)$ is the Jacobson radical.

11.31 Let R^*/R be an integral extension. If $\mathfrak{p}_1 \subseteq \mathfrak{p}_2 \subseteq \cdots \subseteq \mathfrak{p}_n$ is a chain of prime ideals in R and if $\mathfrak{p}_1^* \subseteq \mathfrak{p}_2^* \subseteq \cdots \subseteq \mathfrak{p}_m^*$, for $m < n$ is a chain of prime ideals in R^* with each $\mathfrak{p}_i^*$ lying over $\mathfrak{p}_i$, then the chain in R^* can be extended to $\mathfrak{p}_1^* \subseteq \mathfrak{p}_2^* \subseteq \cdots \subseteq \mathfrak{p}_n^*$ with $\mathfrak{p}_i^*$ lying over $\mathfrak{p}_i$ for all

$i \leq n$. (The going up theorem is so called because the chain of ideals in R^* is ascending, in contrast to the going down theorem in which the chain of ideals in R^* is descending.)

11.32 Let R^*/R be an integral extension. If every nonzero prime ideal in R is a maximal ideal, prove that every nonzero prime ideal in R^* is also a maximal ideal.
Hint. See the proof of Corollary 11.53.

11.33 Let α be algebraic over $\mathbb{Q}$, let $E/\mathbb{Q}$ be a splitting field, and let $G = \mathrm{Gal}(E/\mathbb{Q})$ be its Galois group.

(i) Prove that if α is integral over $\mathbb{Z}$, then, for all $\sigma \in G$, $\sigma(\alpha)$ is also integral over $\mathbb{Z}$.

(ii) Prove that α is an algebraic integer if and only if $\mathrm{irr}(\alpha, \mathbb{Q}) \in \mathbb{Z}[x]$. Compare this proof with that of Corollary 6.29.

(iii) Let E be an algebraic number field and let $R \subseteq E$ be integrally closed. If $\alpha \in E$, prove that $\mathrm{irr}(\alpha, \mathrm{Frac}(R)) \in R[x]$.
Hint. If $\widehat{E}$ is a Galois extension of $\mathrm{Frac}(R)$ containing α, then $G = \mathrm{Gal}(\widehat{E}/\mathrm{Frac}(R))$ acts transitively on the roots of α.

Nullstellensatz Redux

In this subsection, we will prove the Nullstellensatz for arbitrary algebraically closed fields (recall that our proof in Chapter 6 assumed that k is uncountable). There are different proofs of this result, and we present the proof discovered by O. Goldman, as expounded in Kaplansky, *Commutative Rings*.

Definition. A ring extension A/R is ***finitely generated*** if there is a surjective R-algebra map $\varphi\colon R[x_1, \ldots, x_n] \to A$. If $\varphi(x_i) = a_i$, then we write

$$A = R[a_1, \ldots, a_n].$$

If I is an ideal in a commutative ring R, then the nilpotent elements in R/I arise from elements of $\sqrt{I}$. We now begin working toward a theorem of W. Krull that characterizes the nilpotent elements in a commutative ring, for this will give us information about radicals of ideals.

Lemma 11.57. *Let R be a domain with $F = \mathrm{Frac}(R)$. Then F/R is a finitely generated ring extension if and only if F/R is a simply generated ring extension*; *that is, there is $u \in R$ with $F = R[u^{-1}]$ (and so the localization $\{u\}^{-1}R$ is a field).*

Proof. Sufficiency being obvious, we prove only necessity. If $F = R[a_1/b_1, \ldots, a_n/b_n]$, define $u = \prod_i b_i$. We claim that $F = R[u^{-1}]$. Clearly, $F \supseteq R[u^{-1}]$. For the reverse inclusion, note that $a_i/b_i = a_i\widehat{u}_i/u \in R[u^{-1}]$, where $\widehat{u}_i = b_1 \cdots \widehat{b_i} \cdots b_n$. •

Definition. If R is a domain with $F = \mathrm{Frac}(R)$, then R is a ***G-domain*** if F/R is a finitely generated ring extension. An ideal I in a commutative ring R is a ***G-ideal***[4] if R/I is a G-domain.

[4] G-ideals are named after O. Goldman.

Every field is a G-domain, and so every maximal ideal in a commutative ring is a G-ideal. If I is a G-ideal, then R/I is a G-domain, hence a domain; therefore, every G-ideal is a prime ideal. Corollary 11.61 says that $\mathbb{Z}$ is not a G-domain; it follows that the prime ideal (x) in $\mathbb{Z}[x]$ is not a G-ideal.

Proposition 11.58. *Let E/R be a ring extension in which both E and R are domains. If E is a finitely generated R-algebra and each $\alpha \in E$ is algebraic over R (that is, α is a root of a nonzero polynomial in $R[x]$), then R is a G-domain if and only if E is a G-domain.*

Proof. Let R be a G-domain, so that $F = \mathrm{Frac}(R) = R[u^{-1}]$ for some nonzero $u \in R$. Therefore, $E[u^{-1}] \subseteq \mathrm{Frac}(E)$, for $u \in R \subseteq E$. But $E[u^{-1}]$ is a domain algebraic over the field $F = R[u^{-1}]$, so that $E[u^{-1}]$ is a field, by Exercise 11.35 on page 938. Since $\mathrm{Frac}(E)$ is the smallest field containing E, we have $E[u^{-1}] = \mathrm{Frac}(E)$, and so E is a G-domain.

If E is a G-domain, then there is $v \in E$ with $\mathrm{Frac}(E) = E[v^{-1}]$. By hypothesis, $E = R[b_1, \ldots, b_n]$, where b_i is algebraic over $F = \mathrm{Frac}(R)$ for all i. As $v \in E$, we have v algebraic over R, and so v^{-1} is algebraic over R. Thus, there are monic polynomials $f_0(x)$, $f_i(x) \in F[x]$ with $f_0(v^{-1}) = 0$ and $f_i(b_i) = 0$ for all $i \geq 1$. Clearing denominators, we obtain equations $\beta_i f_i(b_i) = 0$, for $i \geq 0$, with coefficients in R:

$$\beta_0(v^{-1})^{d_0} + \cdots = 0$$
$$\beta_i b_i^{d_i} + \cdots = 0.$$

Define $R^* = R[\beta_0^{-1}, \beta_1^{-1}, \ldots, \beta_n^{-1}]$. Each b_i is integral over R^*, for each β_i is a unit in R^*. Clearly, $E[v^{-1}] = R^*[v^{-1}, b_1, \ldots, b_n]$. Thus, the field $E[v^{-1}]$ is integral over R^*, by Proposition 11.43 (since $E[v^{-1}] = R^*[v^{-1}, b_1, \ldots, b_n]$ and each of the displayed generators is integral over R^*), and this forces R^* to be a field, by Proposition 11.51. But $R^* = R[\beta_0^{-1}, \beta_1^{-1}, \ldots, \beta_n^{-1}] \subseteq F$, because $\beta_i \in R$ for all i, so that $R^* = F$. Therefore, $F = R[\beta_0^{-1}, \beta_1^{-1}, \ldots, \beta_n^{-1}]$ is a finitely generated ring extension of R; that is, R is a G-domain. •

The next lemma leads to Corollary 11.60, an "internal" characterization of G-domains, phrased solely in terms of R, with no mention of $\mathrm{Frac}(R)$.

Lemma 11.59. *Let R be a domain with $F = \mathrm{Frac}(R)$. The following conditions are equivalent for a nonzero element $u \in R$.*

(i) *u lies in every nonzero prime ideal of R.*

(ii) *for every nonzero ideal I in R, there is an integer n with $u^n \in I$.*

(iii) *R is a G-domain; that is, $F = R[u^{-1}]$.*

Proof. (i) ⇒ (ii). Suppose there is a nonzero ideal I for which $u^n \notin I$ for all $n \geq 0$. If $S = \{u^n : n \geq 0\}$, then $I \cap S = \varnothing$. By Zorn's lemma, there is an ideal $\mathfrak{p}$ maximal with $I \subseteq \mathfrak{p}$ and $\mathfrak{p} \cap S = \varnothing$, and $\mathfrak{p}$ is a prime ideal, by Exercise 6.9. This contradicts u lying in every prime ideal.

(ii) $\Rightarrow$ (iii). If $b \in R$ and $b \neq 0$, then $u^n \in (b)$ for some $n \geq 1$, by hypothesis. Hence, $u^n = rb$ for some $r \in R$, and so $b^{-1} = ru^{-n} \in R[u^{-1}]$. Therefore,$F = R[u^{-1}]$.

(iii) $\Rightarrow$ (i). Let $\mathfrak{p}$ be a nonzero prime ideal. If $b \in \mathfrak{p}$ is nonzero, then $b^{-1} = \sum_{i=0}^{n} r_i u^{-i}$, where $r_i \in R$, because $F = R[u^{-1}]$. Therefore, $u^n = b\left(\sum_i r_i u^{n-i}\right)$ lies in $\mathfrak{p}$, because $b \in \mathfrak{p}$ and $\sum_i r_i u^{n-i} \in R$. Since $\mathfrak{p}$ is a prime ideal, $u \in \mathfrak{p}$. •

Corollary 11.60. *A domain R is a G-domain if and only if* $\bigcap_{\substack{\mathfrak{p} \text{ prime} \\ \mathfrak{p} \neq 0}} \mathfrak{p} \neq \{0\}$.

Proof. By Lemma 11.59, R is a G-domain if and only if it has a nonzero element u lying in every nonzero prime ideal. •

Corollary 11.61. *If R is a* PID*, then R is a G-domain if and only if R has only finitely many prime ideals.*

Proof. If R is a G-domain, then $I = \bigcap \mathfrak{p} \neq \{0\}$, where $\mathfrak{p}$ ranges over all nonzero prime ideals. If R has infinitely many prime ideals, then there are infinitely many nonassociate *prime elements* $p_1, p_2, \ldots$; that is, the (p_i) are distinct prime ideals. If $a \in I$, then $p_i \mid a$ for all i. But $a = q_1^{e_1} \cdots q_n^{e_n}$, where the q_j are distinct prime elements, contradicting unique factorization in the PID R.

Conversely, if R has only finitely many nonzero prime ideals, say, $(p_1), \ldots, (p_m)$, then the product $p_1 \cdots p_m$ is a nonzero element lying in $\bigcap_i (p_i)$. Therefore, R is a G-domain. •

It follows, for example, that every DVR is a G-domain.

Definition. If R is a commutative ring, then its ***nilradical*** is

$$\text{nil}(R) = \{r \in R : r \text{ is nilpotent}\}.$$

We note that nil(R) is an ideal. If $r, s \in R$ are nilpotent, then $r^n = 0 = s^m$, for positive integers m and n. Hence,

$$(r+s)^{m+n-1} = \sum_{i=0}^{m+n-1} \binom{m+n-1}{i} r^i s^{m+n-1-i}.$$

If $i \geq n$, then $r^i = 0$ and the ith term in the sum is 0; if $i < n$, then $m + n - i - 1 \geq m$, $s^{m+n--1i} = 0$, and the ith term in the sum is 0 in this case as well. Thus, $(r+s)^{m+n-1} = 0$ and $r + s$ is nilpotent. Finally, rs is nilpotent, for $(rs)^{mn} = r^{mn}s^{ms} = 0$.

The next theorem is an improvement on W. Krull's original version, that characterizes the nilradical as the intersection of all the prime ideals.

Theorem 11.62 (Krull). *If R is a commutative ring, then*

$$\text{nil}(R) = \bigcap_{\mathfrak{p}=\text{prime ideal}} \mathfrak{p} = \bigcap_{\mathfrak{p}=G\text{-ideal}} \mathfrak{p}.$$

Remark. If R is a domain, then $\{0\}$ is a prime ideal, and so $\mathrm{nil}(R) = \{0\}$ (alternatively, there are no nonzero nilpotent elements in a domain). The intersection of all the nonzero prime ideals in a commutative ring R may be larger than $\mathrm{nil}(R)$; this happens, for example, when R is a DVR. ◀

Proof. It is obvious that $\mathrm{nil}(R) \subseteq \bigcap_{\mathfrak{p}=\text{prime ideal}} \mathfrak{p} \subseteq \bigcap_{\mathfrak{p}=G\text{-ideal}} \mathfrak{p}$: nilpotent elements lie in every prime ideal; every G-ideal is a prime ideal. Thus, it suffices to prove that $\bigcap_{\mathfrak{p}=G\text{-ideal}} \mathfrak{p} \subseteq \mathrm{nil}(R)$. Suppose that $u \notin \bigcap_{\mathfrak{p}=G\text{-ideal}} \mathfrak{p}$. It follows that $u^n \neq 0$ for all $n \geq 1$, for if $u^n = 0$, then $u^n \in \mathfrak{p}$ for every G-ideal $\mathfrak{p}$, and so $u \in \mathfrak{p}$, because G-ideals are prime. Therefore, the multiplicatively closed set $S = \{u^n : n \geq 1\}$ does not contain 0. By Zorn's lemma, there is an ideal $\mathfrak{q}$ maximal with $\mathfrak{q} \cap S = \varnothing$ (that there are ideals disjoint from S requires $0 \notin S$). We claim that $\mathfrak{q}$ is a G-ideal, which will give $u \notin \bigcap_{\mathfrak{p}=G\text{-ideal}} \mathfrak{p}$. Now $\mathfrak{q}$ is a prime ideal, by Exercise 6.9, so that $R/\mathfrak{q}$ is a domain. Suppose there is a nonzero prime ideal $\mathfrak{p}^*$ in $R/\mathfrak{q}$ not containing $u + \mathfrak{q}$. There is an ideal $\mathfrak{p} \supsetneq \mathfrak{q}$ in R with $\mathfrak{p}^* = \mathfrak{p}/\mathfrak{q}$ (for $\mathfrak{p}^* \neq \{0\}$), contradicting the maximality of $\mathfrak{q}$. Therefore, $u + \mathfrak{q}$ lies in every nonzero prime ideal in $R/\mathfrak{q}$. By Corollary 11.60, $R/\mathfrak{q}$ is a G-domain, and so $\mathfrak{q}$ is a G-ideal. •

The next corollary follows easily from Krull's theorem.

Corollary 11.63. *If I is an ideal in a commutative ring R, then $\sqrt{I}$ is the intersection of all the G-ideals containing I.*

Proof. By definition, $\sqrt{I} = \{r \in R : r^n \in I \text{ for some } n \geq 1\}$. Therefore, $\sqrt{I}/I = \mathrm{nil}(R/I) = \bigcap_{\mathfrak{p}^*=G\text{-ideal}} \mathfrak{p}^*$. For each $\mathfrak{p}^*$, there is an ideal $\mathfrak{p}$ containing I with $\mathfrak{p}^* = \mathfrak{p}/I$, and $\sqrt{I} = \bigcap_{\mathfrak{p}/I=G\text{-ideal}} \mathfrak{p}$. Finally, every $\mathfrak{p}$ involved in the intersection is a G-ideal, because $R/\mathfrak{p} \cong (R/I)/(\mathfrak{p}/I) = (R/I)/\mathfrak{p}^*$ and $(R/I)/\mathfrak{p}^*$ is a G-domain. •

We now focus on the relation between ideals in $R[x]$ and those in R.

Proposition 11.64. *An ideal I in a commutative ring R is a G-ideal if and only if I is the contraction of a maximal ideal in $R[x]$.*

Proof. If I is a G-ideal in R, then R/I is a G-domain. Hence, there is $u \in \mathrm{Frac}(R/I)$ with $\mathrm{Frac}(R/I) = (R/I)[u^{-1}]$. Let $\varphi \colon (R/I)[x] \to (R/I)[u^{-1}]$ be the R-algebra map taking $x \mapsto u^{-1}$. Since φ is a surjection onto the field $(R/I)[u^{-1}] = \mathrm{Frac}(R/I)$, its kernel $\mathfrak{m}$ is a maximal ideal in $(R/I)[x]$. Since $\varphi|(R/I)$ is an injection, we have $\mathfrak{m} \cap (R/I) = \{0\}$. By Exercise 6.2, there is an ideal, necessarily maximal, $\mathfrak{m}'$ in $R[x]$ with $\mathfrak{m}'/I = \mathfrak{m}$, and $\mathfrak{m}' \cap R = I$.

Conversely, assume that $\mathfrak{m}$ is a maximal ideal in $R[x]$ with $\mathfrak{m} \cap R = I$. If $\nu \colon R[x] \to R[x]/\mathfrak{m}$ be the natural map and $u = \nu(x)$, then $\mathrm{im}\, \nu = (R/I)[u]$ is a field. Hence, R/I is a G-domain, by Proposition 11.58, and so I is a G-ideal. •

Notation. If I is an ideal in a commutative ring R and if $f(x) \in R[x]$, then $\overline{f}(x)$ denotes the polynomial in $(R/I)[x]$ obtained from $f(x)$ by reducing its coefficients $\bmod I$; that is, if $f(x) = \sum_i a_i x^i$, where $a_i \in R$, then $\overline{f}(x) = \sum_i (a_i + I)x^i$.

Corollary 11.65. *Let R be a commutative ring, and let $\mathfrak{m}$ be a maximal ideal in $R[x]$. If the contraction $\mathfrak{m}' = \mathfrak{m} \cap R$ is a maximal ideal in R, then $\mathfrak{m} = (\mathfrak{m}', f(x))$, where $f(x) \in R[x]$ and $\overline{f}(x) \in (R/\mathfrak{m}')[x]$ is irreducible. If $R/\mathfrak{m}'$ is algebraically closed, then $\mathfrak{m} = (\mathfrak{m}', x - a)$ for some $a \in R$.*

Proof. First, Proposition 11.64 says that $\mathfrak{m}' = \mathfrak{m} \cap R$ is a G-ideal in R. Consider the map $\varphi\colon R[x] \to (R/\mathfrak{m}')[x]$ which reduces coefficients mod $\mathfrak{m}'$. Since φ is a surjection, the ideal $\varphi(\mathfrak{m})$ is a maximal ideal; that is, $\varphi(\mathfrak{m}) = (g(x))$, where $g(x) \in (R/\mathfrak{m}')[x]$ is irreducible. Therefore, $\mathfrak{m} = (\mathfrak{m}', f(x))$, where $\varphi(f) = g$; that is, $\overline{f}(x) = g(x)$. •

Maximal ideals are always G-ideals, and G-ideals are always prime ideals. The next definition imposes the condition that G-ideals be maximal. In light of Proposition 11.64, this will force the contraction of maximal ideals in $R[x]$ to be maximal ideals in R.

Definition. A commutative ring R is a ***Jacobson***[5] ***ring*** if every G-ideal is a maximal ideal.

Example 11.66.

(i) Every field is a Jacobson ring.

(ii) By Corollary 11.61, a PID R is a G-domain if and only if it has only finitely many prime ideals. Such a ring cannot be a Jacobson ring, for $\{0\}$ is a G-ideal which is not maximal [$R/\{0\} \cong R$ is a G-domain]. On the other hand, if R has infinitely many prime ideals, then R is not a G-domain and $\{0\}$ is not a G-ideal. The G-ideals, which are now nonzero prime ideals, must be maximal. Therefore, a PID is a Jacobson ring if and only if it has infinitely many prime ideals.

(iii) We note that if R is a Jacobson ring, then so is any quotient $R^* = R/I$. If $\mathfrak{p}^*$ is a G-ideal in R^*, then $R^*/\mathfrak{p}^*$ is a G-domain. Now $\mathfrak{p}^* = \mathfrak{p}/I$ for some ideal $\mathfrak{p}$ in R, and $R/\mathfrak{p} \cong (R/I)/(\mathfrak{p}/I) = R^*/\mathfrak{p}^*$. Thus, $\mathfrak{p}$ is a G-ideal in R. Since R is a Jacobson ring, $\mathfrak{p}$ is a maximal ideal, and $R/\mathfrak{p} \cong R^*/\mathfrak{p}^*$ is a field. Therefore, $\mathfrak{p}^*$ is a maximal ideal, and so R^* is also a Jacobson ring.

(iv) By Corollary 11.63, every radical ideal in a commutative ring R is the intersection of all the G-ideals containing it. Therefore, if R is a Jacobson ring, then every radical ideal is an intersection of maximal ideals. ◂

Example 11.66(iv) suggests the following result.

Proposition 11.67. *A commutative ring R is a Jacobson ring if and only if every prime ideal in R is an intersection of maximal ideals.*

Proof. By Corollary 11.63, every radical ideal, hence, every prime ideal, is the intersection of all the G-ideals containing I. But in a Jacobson ring, every G-ideal is maximal.

[5] These rings are called ***Hilbert rings*** by some authors. In 1951, W. Krull and O. Goldman, independently, published proofs of the Nullstellensatz using the techniques in this subsection. Krull introduced the term *Jacobson ring* in his paper.

Conversely, assume that every prime ideal in R is an intersection of maximal ideals. We let the reader check that this property is inherited by quotient rings. Let $\mathfrak{p}$ be a G-ideal in R, so that $R/\mathfrak{p}$ is a G-domain. Thus, there is $u \neq 0$ in $R/\mathfrak{p}$ with $\text{Frac}(R/\mathfrak{p}) = (R/\mathfrak{p})[u^{-1}]$. By Lemma 11.59, u lies in every nonzero prime ideal of $R/\mathfrak{p}$, and so u lies in every nonzero maximal ideal. Now every prime ideal in $R/\mathfrak{p}$ is an intersection of maximal ideals; in particular, since $R/\mathfrak{p}$ is a domain, there are maximal ideals $\mathfrak{m}_\alpha$ with $\{0\} = \cap_\alpha \mathfrak{m}_\alpha$. If all these $\mathfrak{m}_\alpha$ are nonzero, then $u \in \cap_\alpha \mathfrak{m}_\alpha = \{0\}$, a contradiction. We conclude that $\{0\}$ is a maximal ideal. Therefore, $R/\mathfrak{p}$ is a field, the G-ideal $\mathfrak{p}$ is maximal, and R is a Jacobson ring. •

Corollary 11.68. *A commutative ring R is a Jacobson ring if and only if $J(R/I) = \text{nil}(R/I)$ for every ideal I. In particular, $J(R) = \text{nil}(R)$.*

Proof. Let R be a Jacobson ring. If I is an ideal in R, then $\sqrt{I} = \bigcap \mathfrak{m}$, where $\mathfrak{m}$ is a maximal ideal containing I. Now $J(R/I)$ is the intersection of all the maximal ideals in R/I; that is, $J(R) = \bigcap(\mathfrak{m}/I) = (\bigcap \mathfrak{m})/I = \sqrt{I}/I$. On the other hand, $\text{nil}(R/I)$ consists of all the nilpotent elements in R/I. But $0 = (f + I)^n = f^n + I$ holds if and only if $f^n \in I$; that is, $f \in \sqrt{I}$. To prove the converse, note that condition says that every radical ideal in R is an intersection of maximal ideals. In particular, every prime ideal is such an intersection, and so R is a Jacobson ring. •

The next result will give many examples of Jacobson rings.

Theorem 11.69. *A commutative ring R is a Jacobson ring if and only if $R[x]$ is a Jacobson ring.*

Proof. We have seen that every quotient of a Jacobson ring is a Jacobson ring. Hence, if $R[x]$ is a Jacobson ring, then $R \cong R[x]/(x)$ is also a Jacobson ring.

Conversely, suppose that R is a Jacobson ring. If $\mathfrak{q}$ is a G-ideal in $R[x]$, then we may assume that $\mathfrak{q} \cap R = \{0\}$, by Exercise 11.36 on page 938. If $\nu \colon R[x] \to R[x]/\mathfrak{q}$ is the natural map, then $R[x]/\mathfrak{q} = R[u]$, where $u = \nu(x)$. Now $R[u]$ is a G-domain, because $\mathfrak{q}$ is a G-ideal; hence, if $K = \text{Frac}(R[u])$, then there is $v \in K$ with $K = R[u][v^{-1}]$. If $\text{Frac}(R) = F$, then

$$K = R[u][v^{-1}] \subseteq F[u][v^{-1}] \subseteq K,$$

so that $F[u][v^{-1}] = K$; that is, $F[u]$ is a G-domain. But $F[u]$ is not a G-domain if u is transcendental over F, by Corollary 11.61, for $F[x] \cong F[u]$ has infinitely many prime ideals. Thus, u is algebraic over F, and hence u is algebraic over R. Since $R[u]$ is a G-domain, Proposition 11.58 says that R is a G-domain. Now R is a Jacobson ring, and so R is a field, by Exercise 11.34 on page 938. But if R is a field, so is $R[u]$, for u is algebraic over R. Therefore, $R[u] = R[x]/\mathfrak{q}$ is a field, so that $\mathfrak{q}$ is a maximal ideal, and $R[x]$ is a Jacobson ring. •

Corollary 11.70. *If k is a field, then $k[x_1, \ldots, x_n]$ is a Jacobson ring.*

Proof. The proof is by induction on $n \geq 1$. For the base step, $k[x]$ is a PID having infinitely many prime ideals, by Exercise 11.40, and so it is a Jacobson ring, by Example 11.66(ii). For the inductive step, the inductive hypothesis gives $R = k[x_1, \ldots, x_{n-1}]$ a Jacobson ring, and Theorem 11.69 applies. •

Theorem 11.71. *If $\mathfrak{m}$ is a maximal ideal in $k[x_1, \ldots, x_n]$, where k is an algebraically closed field, then there are $a_1, \ldots, a_n \in k$ such that*

$$\mathfrak{m} = (x_1 - a_1, \ldots, x_n - a_n).$$

Proof. The proof is by induction on $n \geq 1$. If $n = 1$, then $\mathfrak{m} = (p(x))$, where $p(x) \in k[x]$ is irreducible. Since k is algebraically closed, $p(x)$ is linear. For the inductive step, let $R = k[x_1, \ldots, x_{n-1}]$. Corollary 11.70 says that R is a Jacobson ring, and so $\mathfrak{m} \cap R$ is a G-ideal in R, by Proposition 11.64. Since R is a Jacobson ring, $\mathfrak{m}' = \mathfrak{m} \cap R$ is a maximal ideal. Corollary 11.65 now applies to give $\mathfrak{m} = (\mathfrak{m}', f(x_n))$, where $f(x_n) \in R[x_n]$ and $\overline{f}(x_n) \in (R/\mathfrak{m}')[x_n]$ is irreducible. As k is algebraically closed and $R/\mathfrak{m}'$ is a finitely generated k-algebra, $R/\mathfrak{m}' \cong k$, and we may assume that $f(x_n)$ is linear; there is $a_n \in k$ with $f_n(x) = x_n - a_n$. By the inductive hypothesis, $\mathfrak{m}' = (x_1 - a_1, \ldots, x_{n-1} - a_{n-1})$ for $a_1, \ldots, a_{n-1} \in k$, and this completes the proof. •

We use Theorem 11.71 to prove the Weak Nullstellensatz, Theorem 6.100. Recall that only the special case of the Nullstellensatz for uncountable algebraically closed fields was proved in Chapter 6.

Theorem 11.72 (Weak Nullstellensatz). *If $f_1(X), \ldots, f_t(X) \in k[X]$, where k is an algebraically closed field, then $I = (f_1, \ldots, f_t)$ is a proper ideal in $k[X]$ if and only if* $\operatorname{Var}(f_1, \ldots, f_t) \neq \varnothing$.

Proof. If I is a proper ideal, then there is a maximal ideal $\mathfrak{m}$ containing it. By Theorem 6.100, there is $a = (a_1, \ldots, a_n) \in k^n$ with $\mathfrak{m} = (x_1 - a_1, \ldots, x_n - a_n)$. Now $I \subseteq \mathfrak{m}$ implies $\operatorname{Var}(\mathfrak{m}) \subseteq \operatorname{Var}(I)$. But $a \in \operatorname{Var}(\mathfrak{m})$, and so $\operatorname{Var}(I) \neq \varnothing$. •

We could now repeat the proof of the Nullstellensatz over $\mathbb{C}$, Theorem 6.102, to obtain the Nullstellensatz over any algebraically closed field. However, the following proof is easier.

Theorem 11.73 (Nullstellensatz). *Let k be an algebraically closed field. If I is an ideal in $k[x_1, \ldots, x_n]$, then* $\operatorname{Id}(\operatorname{Var}(I)) = \sqrt{I}$.

Proof. The inclusion $\operatorname{Id}(\operatorname{Var}(I)) \supseteq \sqrt{I}$ is easy to see. If $f^n(a) = 0$ for all $a \in \operatorname{Var}(I)$, then $f(a) = 0$ for all $a \in \operatorname{Var}(I)$, because the values of f lie in the field k. Hence, $f \in \operatorname{Id}(\operatorname{Var}(I))$. For the reverse inclusion, note first that $k[x_1, \ldots, x_n]$ is a Jacobson ring, by Corollary 11.70; hence, Example 11.66(iv) shows that $\sqrt{I}$ is an intersection of maximal ideals. Let $g \in \operatorname{Id}(\operatorname{Var}(I))$. If $\mathfrak{m}$ is a maximal ideal containing I, then $\operatorname{Var}(\mathfrak{m}) \subseteq \operatorname{Var}(I)$, and so $\operatorname{Id}(\operatorname{Var}(I)) \subseteq \operatorname{Id}(\operatorname{Var}(\mathfrak{m}))$. But $\operatorname{Id}(\operatorname{Var}(\mathfrak{m})) = \mathfrak{m}$: $\operatorname{Id}(\operatorname{Var}(I)) \supseteq \sqrt{\mathfrak{m}} = \mathfrak{m}$, because $\mathfrak{m}$ is a maximal, hence prime ideal. Therefore, $g \in \bigcap \mathfrak{m} = \sqrt{I}$, as desired. •

EXERCISES

11.34 Prove that a commutative ring R is a field if and only if R is a Jacobson ring and a G-domain.

11.35 Let E/R be a ring extension in which R is a field and E is a domain.

(i) Let $b \in E$ be algebraic over R, prove that there exists an equation

$$b^n + r_{n-1}b^{n-1} + \cdots + r_1 b + r_0 = 0,$$

where $r_i \in R$ for all i and $r_0 \neq 0$.

(ii) If $E = R[b_1, \ldots, b_m]$, where each b_j is algebraic over R, prove that E is a field.

11.36 Let R be a Jacobson ring, and assume that $(R/\mathfrak{q}')[x]$ is a Jacobson ring for every G-ideal $\mathfrak{q}$ in $R[x]$, where $\mathfrak{q}' = \mathfrak{q} \cap R$. Prove that $R[x]$ is a Jacobson ring.

11.37 (i) Prove that $\mathfrak{m} = (x^2 - y, y^2 - 2)$ is a maximal ideal in $\mathbb{Q}[x, y]$.

(ii) Prove that there do not exist $f(x) \in \mathbb{Q}[x]$ and $g(y) \in \mathbb{Q}[y]$ with $\mathfrak{m} = \big(f(x), g(y)\big)$.

11.38 Let k be a field and let $\mathfrak{m}$ be a maximal ideal in $k[x_1, \ldots, x_n]$. Prove that

$$\mathfrak{m} = \big(f_1(x_1),\ f_2(x_1, x_2), \ldots,\ f_{n-1}(x_1, \ldots, x_{n-1}),\ f_n(x_1, \ldots, x_n)\big).$$

Hint. Use Corollary 11.65.

11.39 Prove that if R is noetherian, then $\mathrm{nil}(R)$ is a nilpotent ideal.

11.40 If k is a field, prove that $k[x]$ has infinitely many prime ideals.

Algebraic Integers

We have mentioned that Kummer investigated the ring $\mathbb{Z}[\zeta_p]$, where p is an odd prime and ζ_p is a primitive pth root of unity. We now study rings of integers in algebraic number fields E further. Recall the definition:

$$\mathcal{O}_E = \{\alpha \in E : \alpha \text{ is integral over } \mathbb{Z}\}.$$

We begin with a consequence of Gauss's lemma.

Lemma 11.74. *Let E be an algebraic number field with $[E : \mathbb{Q}] = n$, and let $\alpha \in E$ be an algebraic integer. Then* $\mathrm{irr}(\alpha, \mathbb{Q}) \in \mathbb{Z}[x]$ *and* $\deg(\mathrm{irr}(\alpha, \mathbb{Q})) \mid n$.

Proof. By Corollary 6.29, $\mathrm{irr}(\alpha, \mathbb{Q}) \in \mathbb{Z}[x]$, and so the result follows from Proposition 3.117(v). •

Definition. A ***quadratic field*** is an algebraic number field E with $[E : \mathbb{Q}] = 2$.

Proposition 11.75. *Every quadratic field E has the form $E = \mathbb{Q}(\sqrt{d})$, where d is a squarefree integer.*

Proof. We know that $E = \mathbb{Q}(\alpha)$, where α is a root of a quadratic polynomial; say, $\alpha^2 + b\alpha + c = 0$, where $b, c \in \mathbb{Q}$. If $D = b^2 - 4c$, then the quadratic formula gives $\alpha = -\frac{1}{2}b \pm \frac{1}{2}\sqrt{D}$, and so $E = \mathbb{Q}(\alpha) = \mathbb{Q}(\sqrt{D})$. Write D in lowest terms: $D = U/V$, where $U, V \in \mathbb{Z}$ and $(U, V) = 1$. Now $U = ur^2$ and $V = vs^2$, where u, v are squarefree; hence, uv is squarefree, because $(u, v) = 1$. Therefore, $\mathbb{Q}(\sqrt{D}) = \mathbb{Q}(\sqrt{u/v}) = \mathbb{Q}(\sqrt{uv})$, for $\sqrt{u/v} = \sqrt{uv/v^2} = \sqrt{uv}/v$. •

We now describe the integers in quadratic fields.

Proposition 11.76. *Let $E = \mathbb{Q}(\sqrt{d})$, where d is a squarefree integer (which implies that $d \not\equiv 0 \bmod 4$).*

(i) *If $d \equiv 2 \bmod 4$ or $d \equiv 3 \bmod 4$, then $\mathcal{O}_E = \mathbb{Z}[\sqrt{d}]$.*

(ii) *If $d \equiv 1 \bmod 4$, then $\mathcal{O}_E$ consists of all $\frac{1}{2}\big(u + v\sqrt{d}\big)$ with u and v rational integers having the same parity.*

Proof. If $\alpha \in E = \mathbb{Q}(\sqrt{d})$, then there are $a, b \in \mathbb{Q}$ with $\alpha = a + b\sqrt{d}$. We first show that $\alpha \in \mathcal{O}_E$ if and only if

$$2a \in \mathbb{Z} \qquad \text{and} \qquad a^2 - db^2 \in \mathbb{Z}. \tag{3}$$

If $\alpha \in \mathcal{O}_E$, then Lemma 11.74 says that $p(x) = \operatorname{irr}(\alpha, \mathbb{Q}) \in \mathbb{Z}[x]$ is quadratic. Now $\operatorname{Gal}(E/\mathbb{Q}) = \langle\sigma\rangle$, where $\sigma : E \to E$ carries $\sqrt{d} \mapsto -\sqrt{d}$; that is,

$$\sigma(\alpha) = a - b\sqrt{d}.$$

Since σ permutes the roots of $p(x)$, the other root of $p(x)$ is $\sigma(\alpha)$; that is,

$$p(x) = (x - \alpha)(x - \sigma(\alpha)) = x^2 - 2ax + (a^2 - db^2).$$

Hence, Eqs. (3) hold, because $p(x) \in \mathbb{Z}[x]$.

Conversely, if Eqs. (3) hold, then $\alpha \in \mathcal{O}_E$, because α is a root of $x^2 - 2ax + (a^2 - db^2)$, a monic polynomial in $\mathbb{Z}[x]$.

We now show that $2b \in \mathbb{Z}$. Multiplying the second equation in (3) by 4 gives $(2a)^2 - d(2b)^2 \in \mathbb{Z}$. Since $2a \in \mathbb{Z}$, we have $d(2b)^2 \in \mathbb{Z}$. Write $2b$ in lowest terms: $2b = m/n$, where $(m, n) = 1$. Now $dm^2/n^2 \in \mathbb{Z}$, so that $n^2 \mid dm^2$. But $(n^2, m^2) = 1$ forces $n^2 \mid d$; as d is squarefree, $n = 1$ and $2b = m/n \in \mathbb{Z}$.

We have shown that $a = \frac{1}{2}u$ and $b = \frac{1}{2}v$, where $u, v \in \mathbb{Z}$. Substituting these values into the second equation in (3) gives

$$u^2 \equiv dv^2 \bmod 4. \tag{4}$$

Note that squares are congruent, mod 4, either to 0 or to 1. If $d \equiv 2 \bmod 4$, then the only way to satisfy Eq. (4) is $u^2 \equiv 0 \bmod 4$ and $v^2 \equiv 0 \bmod 4$. Thus, both u and v must

be even, and so $\alpha = \frac{1}{2}u + \frac{1}{2}v\sqrt{d} \in \mathbb{Z}[\sqrt{d}]$. Therefore, $\mathcal{O}_E = \mathbb{Z}[\sqrt{d}]$ in this case, for $\mathbb{Z}[\sqrt{d}] \subseteq \mathcal{O}_E$ is easily seen to be true. A similar argument works when $d \equiv 3 \bmod 4$. However, if $d \equiv 1 \bmod 4$, then $u^2 \equiv v^2 \bmod 4$. Hence, v is even if and only if u is even; that is, u and v have the same parity. If u and v are both even, then $a, b \in \mathbb{Z}$ and $\alpha \in \mathcal{O}_E$. If u and v are both odd, then $u^2 \equiv 1 \equiv v^2 \bmod 4$, and so $u^2 \equiv dv^2 \bmod 4$, because $d \equiv 1 \bmod 4$. Therefore, Eqs. (3) do hold, and so α lies in $\mathcal{O}_E$. •

If $E = \mathbb{Q}(\sqrt{d})$, where $d \in \mathbb{Z}$, then $\mathbb{Z}[\sqrt{d}] \subseteq \mathcal{O}_E$, but we now see that this inclusion may be strict. For example, $\frac{1}{2}\big(1 + \sqrt{5}\big)$ is an algebraic integer (it is a root of $x^2 - x - 6$). Therefore, $\mathbb{Z}[\sqrt{5}] \subsetneq \mathcal{O}_E$, where $E = \mathbb{Q}(\sqrt{5})$.

The coming brief digression into linear algebra will enable us to prove that rings of integers $\mathcal{O}_E$ are noetherian.

Definition. Let E/k be a field extension in which E is finite-dimensional. If $u \in E$, then multiplication $\Gamma_u \colon E \to E$, given by $\Gamma_u \colon y \mapsto uy$, is a k-map. If $e_1, \ldots, e_n$ is a basis of E, then Γ_u is represented by a matrix $A = [a_{ij}]$ with entries in k; that is,

$$\mu(e_i) = ue_i = \sum a_{ij}e_j.$$

Define the ***trace*** $\mathrm{tr}(u) = \mathrm{tr}(\Gamma_u)$ and the ***norm*** $N(u) = \det(\Gamma_u)$. Define the ***trace form*** $t \colon E \times E \to R$ by

$$t(u, v) = \mathrm{tr}(uv) = \mathrm{tr}(\Gamma_{uv}).$$

The characteristic polynomial of a linear transformation, and hence, any of its coefficients, is independent of any choice of basis of E/k, and so the definitions of trace and norm do not depend on the choice of basis. If $u \in k$, then the matrix of Γ_u, with respect to any basis of E/k, is the scalar matrix uI. Hence,

$$\mathrm{tr}(u) = [E : k]u \quad \text{and} \quad N(u) = u^{[E:k]} \qquad \text{if } u \in k.$$

It is also easy to see that $\mathrm{tr} \colon E \to k$ is a linear functional and that $N \colon E^\times \to k^\times$ is a (multiplicative) homomorphism.

It is a routine exercise, left to the reader, to check that the trace form is a symmetric bilinear form.

Example 11.77.
If $E = \mathbb{Q}(\sqrt{d})$ is a quadratic field, then a basis for $E/\mathbb{Q}$ is $1, \sqrt{d}$. If $u = a + b\sqrt{d}$, then the matrix of Γ_u is

$$\begin{bmatrix} a & bd \\ b & a \end{bmatrix},$$

so that

$$\mathrm{tr}(u) = 2a \qquad \text{and} \qquad N(u) = a^2 - db^2.$$

Thus, trace and norm arose in the description of the integers in quadratic fields, in Eqs. (3).

We now show that $u = a + b\sqrt{d}$ is a unit in $\mathcal{O}_E$ if and only if $N(u) = \pm 1$. If u is a unit, then there is $v \in \mathcal{O}_E$ with $1 = uv$. Hence, $1 = N(1) = N(uv) = N(u)N(v)$, so that $N(u)$ is a unit in $\mathbb{Z}$; that is, $N(u) = \pm 1$. Conversely, if $N(u) = \pm 1$, then $N(\overline{u}) = N(u) = \pm 1$, where $\overline{u} = a - b\sqrt{d}$. Therefore, $N(u\overline{u}) = 1$. But $u\overline{u} \in \mathbb{Q}$, so that $1 = N(u\overline{u}) = (u\overline{u})^2$. Therefore $u\overline{u} = \pm 1$, and so u is a unit. ◂

Lemma 11.78. *Let E/k be a field extension of finite degree n, and let $u \in E$. If $u = u_1, \ldots, u_s$ are the roots, with multiplicity, of* $\operatorname{irr}(u, k)$ *(in some extension field of E), that is,* $\operatorname{irr}(u, k) = \prod_{i=1}^{s}(x - u_i)$*, then*

$$\operatorname{tr}(u) = [E : k(u)] \sum_{i=1}^{s} u_i \qquad \text{and} \qquad N(u) = \Big(\prod_{i=1}^{s} u_i\Big)^{[E:k(u)]}.$$

Remark. Of course, if u is separable over k, then $\operatorname{irr}(u, k)$ has no repeated roots and each u_i occurs exactly once in the formulas. ◂

Sketch of Proof. A basis of $k(u)$ over k is $1, u, u^2, \ldots, u^{s-1}$, and the matrix C_1 of $\Gamma_u | k(u)$ with respect to this basis is the companion matrix of $\operatorname{irr}(u, k)$. If $1, v_2, \ldots, v_r$ is a basis of E over $k(u)$, then the list

$$1, u, \ldots, u^{s-1}, v_1, v_1 u, \ldots, v_1 u^{s-1}, \ldots, v_r, v_r u, \ldots, v_r u^{s-1}$$

is a basis of E over k. Each of the subspaces $k(u)$ and $\langle v_j, v_j u, \ldots, v_j u^{s-1}\rangle$ for $j \geq 2$ is Γ_u-invariant, and so the matrix of Γ_u relative to the displayed basis of E over k is a direct sum of blocks $C_1 \oplus \cdots \oplus C_r$. In fact, the reader may check that each C_j is the companion matrix of $\operatorname{irr}(u, k)$. The trace and norm formulas now follow from $\operatorname{tr}(C_1 \oplus \cdots \oplus C_r) = \sum_j \operatorname{tr}(C_j)$ and $\det(C_1 \oplus \cdots \oplus C_r) = \prod_j \det(C_j)$. •

If E/k is a field extension and $u \in E$, then a more precise notation for the trace and norm is

$$\operatorname{tr}_{E/k}(u) \qquad \text{and} \qquad N_{E/k}(u).$$

Indeed, the formulas in Lemma 11.78 display the dependence on the larger field E.

Proposition 11.79. *Let R be a domain with $F =$* $\operatorname{Frac}(R)$*, let E/F be a field extension of finite degree $[E : F] = n$, and let $u \in E$ be integral over R. If R is integrally closed, then*

$$\operatorname{tr}(u) \in R \qquad \textit{and} \qquad N(u) \in R.$$

Proof. The formulas for $\operatorname{tr}(u)$ and $N(u)$ in Lemma 11.78 express each as an elementary symmetric function of the roots $u = u_1, \ldots, u_s$ of $\operatorname{irr}(u, F)$. Since u is integral over R, Exercise 11.33(iii) on page 931 says that $\operatorname{irr}(u, F) \in R[x]$. Therefore, $\sum_i u_i$ and $\prod_i u_i$ lie in R, and hence $tr(u)$ and $N(u)$ lie in R. •

In Example 4.35, we saw that if E/k is a finite separable extension, then its normal closure $\widehat{E}$ is a Galois extension of k. Recall from the fundamental theorem of Galois theory, Theorem 4.43, that if $G = \text{Gal}(\widehat{E}/k)$ and $H = \text{Gal}(\widehat{E}/E)$, then $[G : H] = [E : k]$.

Lemma 11.80. *Let E/k be a separable field extension of finite degree $n = [E : k]$ and let $\widehat{E}$ be a normal closure of E. Write $G = \text{Gal}(\widehat{E}/k)$ and $H = \text{Gal}(\widehat{E}/E)$, and let T be a transversal of H in G; that is, there is a disjoint union $G = \bigcup_{\sigma \in T} \sigma H$.*

(i) *For all $u \in E$,*

$$\prod_{\sigma \in T} (x - \sigma(u)) = \text{irr}(u, k)^{[E:k(u)]}.$$

(ii) *For all $u \in E$,*

$$\text{tr}(u) = \sum_{\sigma \in T} \sigma(u) \quad \textit{and} \quad N(u) = \prod_{\sigma \in T} \sigma(u).$$

Proof. (i) Denote $\prod_{\sigma \in T}(x - \sigma(u))$ by $h(x)$; of course, $h(x) \in \widehat{E}[x]$.

We claim that the set X, defined by $X = \{\sigma(u) : \sigma \in T\}$, satisfies $\tau(X) = X$ for every $\tau \in G$. If $\sigma \in T$, then $\tau\sigma \in \sigma' H$ for some $\sigma' \in T$, because T is a left transversal; hence, $\tau\sigma = \sigma'\eta$ for some $\eta \in H$. But $\tau\sigma(u) = \sigma'\eta(u) = \sigma'(u)$, because $\eta \in H$, and every element of H fixes E. Therefore, $\tau\sigma(u) = \sigma'(u) \in X$. Thus, the function φ_τ, defined by $\sigma(u) \mapsto \tau\sigma(u)$, is a function $X \to X$. In fact, φ_τ is a permutation, because τ is an isomorphism and so $\varphi_\tau | X$ is an injection. It follows that every elementary symmetric function on $X = \{\sigma(u) : \sigma \in T\}$ is fixed by every $\tau \in G$. Since $\widehat{E}/k$ is a Galois extension, each value of these elementary symmetric functions lies in k. We have shown that all the coefficients of $h(x)$ lie in k, and so $h(x) \in k[x]$. We now compare $h(x)$ and $\text{irr}(u, k)$. If $\sigma \in G$, then σ permutes the roots of $\text{irr}(u, k)$, so that every root $\sigma(u)$ of $h(x)$ is also a root of $\text{irr}(u, k)$. By Exercise 3.86 on page 197, we have

$$h(x) = \text{irr}(u, k)^m$$

for some $m \geq 1$, and so it only remains to compute m. Now

$$\deg(h) = m \deg(\text{irr}(u, k)) = m[k(u) : k].$$

But $\deg(h) = [G : H] = [E : k]$, and so $m = [E : k]/[k(u) : k] = [E : k(u)]$.

(ii) Recall our earlier notation: $\text{irr}(u, k) = \prod_{i=1}^{s}(x - u_i)$. Since

$$\prod_{\sigma \in T} (x - \sigma(u)) = \text{irr}(u, k)^{[E:k(u)]} = \Big(\prod_{i=1}^{s} (x - u_i)\Big)^{[E:k(u)]},$$

their constant terms are the same,

$$\pm \prod_{\sigma \in T} \sigma(u) = \pm \Big(\prod_{i=1}^{s} u_i\Big)^{[E:k(u)]},$$

and their penultimate coefficients are the same,

$$-\sum_{\sigma \in T} \sigma(u) = -[E : k(u)] \sum_{i=1}^{s} u_i.$$

By Lemma 11.78, $\mathrm{tr}(u) = [E : k(u)] \sum_{i=1}^{s} u_i$ and $N(u) = \big(\prod_{i=1}^{s} u_i\big)^{[E:k(u)]}$. It follows that

$$\mathrm{tr}(u) = [E : k(u)] \sum_{i=1}^{s} u_i = \sum_{\sigma \in T} \sigma(u)$$

and

$$N(u) = \big(\prod_{i=1}^{s} u_i\big)^{[E:k(u)]} = \prod_{\sigma \in T} \sigma(u). \quad \bullet$$

Definition. Let E/k be a finite field extension, let $\widehat{E}$ be a normal closure of E, and let T be a left transversal of $\mathrm{Gal}(\widehat{E}/E)$ in $\mathrm{Gal}(\widehat{E}/k)$. If $u \in E$, then the elements $\sigma(u)$, where $\sigma \in T$, are called the ***conjugates*** of u.

If E/k is a separable extension, then the conjugates of u are the roots of $\mathrm{irr}(u, k)$; in the inseparable case, then the conjugates may occur with multiplicities.

Corollary 11.81. *If E/k is a Galois extension with $G = \mathrm{Gal}(E/k)$, then*

$$\mathrm{tr}(u) = \sum_{\sigma \in G} \sigma(u) \quad \textit{and} \quad N(u) = \prod_{\sigma \in G} \sigma(u).$$

Proof. Since E/k is a Galois extension, E is its own normal closure, and so a transversal T of G in itself is just G. $\bullet$

This last corollary shows that the norm here coincides with the norm occurring in Chapter 4 in the proof of Hilbert's Theorem 90.

Let V be a vector space over a field k, and let $f : V \times V \to k$ be a bilinear form. If $e_1, \ldots, e_n$ is a basis of V, then the *discriminant* is defined by

$$D(e_1, \ldots, e_n) = \det([f(e_i, e_j)]).$$

Recall that f is *nondegenerate* if there is a basis whose discriminant is nonzero (it then follows that the discriminant of f with respect to any other basis of V is also nonzero).

Lemma 11.82. *If E/k is a finite separable*[6] *field extension, then the trace form is nondegenerate.*

[6] If E/k is inseparable, then the trace form is identically 0. See Isaacs, *Algebra, A Graduate Course*, page 369.

Proof. We compute the discriminant using Lemma 11.80 (which uses separability). Let $T = \{\sigma_1, \ldots, \sigma_n\}$ be a transversal of $\mathrm{Gal}(\widehat{E}/E)$ in $\mathrm{Gal}(\widehat{E}/k)$, where $\widehat{E}$ is a normal closure of E.

$$\begin{aligned} D(e_1, \ldots, e_n) &= \det([t(e_i, e_j)]) \\ &= \det([\mathrm{tr}(e_i e_j)]) \\ &= \det\big[\sum_\ell \sigma_\ell(e_i e_j)\big] \qquad \text{(Lemma 11.80)} \\ &= \det\big(\big[\sum_\ell \sigma_\ell(e_i)\sigma_\ell(e_j)\big]\big) \\ &= \det([\sigma_\ell(e_i)]) \det([\sigma_\ell(e_j)]) \\ &= \det([\sigma_\ell(e_i)])^2. \end{aligned}$$

To see that $\det([\sigma_\ell(e_i)]) \neq 0$, we assume otherwise. If $[\sigma_\ell(e_i)]$ is singular, there is a column matrix $C = [c_1, \ldots, c_n]^t \in \widehat{E}^n$ with $[\sigma_\ell(e_i)]C = 0$. Hence,

$$c_1\sigma_1(e_j) + \cdots + c_n\sigma_n(e_j) = 0$$

for $j = 1, \ldots, n$. It follows that

$$c_1\sigma_1(v) + \cdots + c_n\sigma_n(v) = 0$$

for every linear combination v of the e_i. But this contradicts the independence of characters, Proposition 4.30. •

Proposition 11.83. *Let R be integrally closed, and let $F = \mathrm{Frac}(R)$. If E/F is a finite separable field extension of degree n, and if $\mathcal{O} = \mathcal{O}_{E/R}$ is the integral closure of R in E, then $\mathcal{O}$ can be imbedded as a submodule of a free R-module of rank n.*

Proof. Let $e_1, \ldots, e_n$ be a basis of E/F. By Proposition 11.46, for each i there is $r_i \in R$ with $r_i e_i \in \mathcal{O}$; changing notation if necessary, we may assume that each $e_i \in \mathcal{O}$. Now Corollary 9.76, which uses nondegeneracy of bilinear forms, says that there is a basis $f_1, \ldots, f_n$ of E with $t(e_i, f_j) = \mathrm{tr}(e_i f_j) = \delta_{ij}$.

Let $\alpha \in \mathcal{O}$. Since $f_1, \ldots, f_n$ is a basis, there are $c_j \in F$ with $\alpha = \sum c_j f_j$. For each i, where $1 \leq i \leq n$, we have $e_i\alpha \in \mathcal{O}$ (because $e_i \in \mathcal{O}$). Therefore, $\mathrm{tr}(e_i\alpha) \in R$, by Proposition 11.79. But

$$\begin{aligned} \mathrm{tr}(e_i\alpha) &= \mathrm{tr}\big(\sum_j c_j e_i f_j\big) \\ &= \sum_j c_j \,\mathrm{tr}(e_i f_j) \\ &= c_j\delta_{ij} \\ &= c_i. \end{aligned}$$

Therefore, $c_i \in R$ for all i, and so $\alpha = \sum_i c_i f_i$ lies in the free R-module with basis $f_1, \ldots, f_n$. •

Definition. If E is an algebraic number field, then an ***integral basis*** for $\mathcal{O}_E$ is a list $\beta_1, \ldots, \beta_n$ in $\mathcal{O}_E$ such that every $\alpha \in \mathcal{O}_E$ has a unique expression

$$\alpha = c_1\beta_1 + \cdots + c_n\beta_n,$$

where $c_i \in \mathbb{Z}$ for all i.

We now prove that integral bases always exist.

Proposition 11.84. *Let E be an algebraic number field.*

(i) *The ring of integers $\mathcal{O}_E$ has an integral basis, and hence it is a free abelian group of finite rank under addition.*

(ii) *$\mathcal{O}_E$ is a noetherian domain.*

Proof. (i) Since $\mathbb{Q}$ has characteristic 0, the field extension $E/\mathbb{Q}$ is separable. Hence, Proposition 11.83 applies to show that $\mathcal{O}_E$ is a submodule of a free $\mathbb{Z}$-module of finite rank; that is, $\mathcal{O}_E$ is a subgroup of a finitely generated free abelian group. By Corollary 9.4, $\mathcal{O}_E$ is itself a free abelian group. But a basis of $\mathcal{O}_E$ as a free abelian group is an integral basis.

(ii) Any ideal I in $\mathcal{O}_E$ is a subgroup of a finitely generated free abelian group, and hence I is itself a finitely generated abelian group, by Proposition 9.7. A fortiori, I is a finitely generated $\mathcal{O}_E$-module; that is, I is a finitely generated ideal. •

Example 11.85.
We show that $\mathcal{O}_E$ need not be a UFD, and hence it need not be a PID. Let $E = \mathbb{Q}(\sqrt{-5})$. Since $-5 \equiv 3 \bmod 4$, Proposition 11.76 gives $\mathcal{O}_E = \mathbb{Z}[\sqrt{-5}]$. By Example 11.77, the only units in $\mathcal{O}_E$ are elements u with $N(u) = \pm 1$. If $a^2 + 5b^2 = \pm 1$, where $a, b \in \mathbb{Z}$, then $b = 0$ and $a = \pm 1$, and so the only units in $\mathcal{O}_E$ are ± 1. Consider the factorization in $\mathcal{O}_E$:

$$2 \cdot 3 = \big(1 + \sqrt{-5}\big)\big(1 - \sqrt{-5}\big).$$

Note that no two of these factors are associates (the only units are ± 1), and we now show that each of them is irreducible. If $v \in \mathcal{O}_E$ divides any of these four factors (but is not an associate of it), then $N(v)$ is a proper divisor in $\mathbb{Z}$ of 4, 9, or 6, for these are the norms of the four factors ($N\big(1+\sqrt{-5}\big) = 6 = N\big(1-\sqrt{-5}\big)$). It is quickly checked, however, that there are no such divisors in $\mathbb{Z}$ of the form $a^2 + 5b^2$ other than ± 1. Therefore, $\mathcal{O}_E = \mathbb{Z}[\sqrt{-5}\,]$ is not a UFD. ◀

Trace and norm can be used to find other rings of integers.

Definition. If $n \geq 2$, then a ***cyclotomic field*** is $E = \mathbb{Q}(\zeta_n)$, where ζ_n is a primitive nth root of unity.

Recall that if p is prime, then the cyclotomic polynomial

$$\Phi_p(x) = x^{p-1} + x^{p-2} + \cdots + x + 1 \in \mathbb{Z}[x]$$

is irreducible, so that $\operatorname{irr}(\zeta_p, \mathbb{Q}) = \Phi_p(x)$ and $[\mathbb{Q}(\zeta_p)/\mathbb{Q}] = p - 1$. Moreover,

$$\operatorname{Gal}(\mathbb{Q}(\zeta_p)/\mathbb{Q}) = \{\sigma_1, \ldots, \sigma_{p-1}\},$$

where $\sigma_i \colon \zeta_p \mapsto \zeta_p^i$ for $i = 1, \ldots, p-1$.

We do some elementary calculations in $E = \mathbb{Q}(\zeta_p)$ to enable us to describe $\mathcal{O}_E$.

Lemma 11.86. *Let p be an odd prime, and let $E = \mathbb{Q}(\zeta)$, where $\zeta = \zeta_p$ is a primitive pth root of unity.*

(i) $\operatorname{tr}(\zeta^i) = -1$ *for* $1 \le i \le p-1$.

(ii) $\operatorname{tr}(1 - \zeta^i) = p$ *for* $1 \le i \le p-1$.

(iii) $p = \prod_{i=1}^{p-1}(1 - \zeta^i) = N(1 - \zeta)$.

(iv) $\mathcal{O}_E(1 - \zeta) \cap \mathbb{Z} = p\mathbb{Z}$.

(v) $\operatorname{tr}(u(1 - \zeta)) \in p\mathbb{Z}$ *for every* $u \in \mathcal{O}_E$.

Proof. (i) We have $\operatorname{tr}(\zeta) = \sum_{i=1}^{p-1} \zeta^i = \Phi_p(\zeta) - 1$, which is also true for every primitive pth root of unity ζ^i. The result follows from $\Phi_p(\zeta) = 0$.

(ii) Since $\operatorname{tr}(1) = [E : \mathbb{Q}] = p - 1$ and tr is a linear functional,

$$\operatorname{tr}(1 - \zeta^i) = \operatorname{tr}(1) - \operatorname{tr}(\zeta^i) = (p - 1) - (-1) = p.$$

(iii) Since $\Phi(x) = x^{p-1} + x^{p-2} + \cdots + x + 1$, we have $\Phi_p(1) = p$. On the other hand, the primitive pth roots of unity are the roots of $\Phi_p(x)$, so that

$$\Phi_p(x) = \prod_{i=1}^{p-1}(x - \zeta^i).$$

Evaluating at $x = 1$ gives the first equation. The second equation holds because the $1 - \zeta^i$s are the conjugates of $1 - \zeta$.

(iv) The first equation in (iii) shows that $p \in \mathcal{O}_E(1 - \zeta) \cap \mathbb{Z}$, so that $\mathcal{O}_E(1 - \zeta) \cap \mathbb{Z} \supseteq p\mathbb{Z}$. If this inclusion is strict, then $\mathcal{O}_E(1 - \zeta) \cap \mathbb{Z} = \mathbb{Z}$, because $p\mathbb{Z}$ is a maximal ideal in $\mathbb{Z}$. In this case, $\mathcal{O}_E(1 - \zeta) \cap \mathbb{Z} = \mathbb{Z}$, hence $\mathbb{Z} \subseteq \mathcal{O}_E(1 - \zeta)$, and so $1 \in \mathcal{O}_E(1 - \zeta)$. Thus, there is $v \in \mathcal{O}_E$ with $v(1 - \zeta) = 1$; that is, $1 - \zeta$ is a unit in $\mathcal{O}_E$. But if $1 - \zeta$ is a unit, then $N(1 - \zeta) = \pm 1$, contradicting the second equation in (iii).

(v) Each conjugate $\sigma_i(u(1 - \zeta)) = \sigma_i(u)(1 - \zeta^i)$ is, obviously, divisible by $1 - \zeta^i$ in $\mathcal{O}_E$. But $1 - \zeta^i$ is divisible by $1 - \zeta$ in $\mathcal{O}_E$, because

$$1 - \zeta^i = (1 - \zeta)(1 + \zeta + \zeta^2 + \cdots + \zeta^{i-1}).$$

Hence, $\sigma_i(1-\zeta^i) \in \mathcal{O}_E(1-\zeta)$ for all i, and so $\sum_i(u(1-\zeta^i)) \in \mathcal{O}_E(1-\zeta)$. By Corollary 11.81, $\sum_i(u(1-\zeta)) = \text{tr}(u(1-\zeta)$. Therefore, $\text{tr}(u(1-\zeta) \in \mathcal{O}_E(1-\zeta) \cap \mathbb{Z} = p\mathbb{Z}$, by (iv), for $\text{tr}(u(1-\zeta)) \in \mathbb{Z}$, by Proposition 11.79. •

Proposition 11.87. *If p is an odd prime and $E = \mathbb{Q}(\zeta_p)$ is a cyclotomic field, then*

$$\mathcal{O}_E = \mathbb{Z}[\zeta_p].$$

Proof. Let us abbreviate ζ_p to ζ. It is always true that $\mathbb{Z}[\zeta] \subseteq \mathcal{O}_E$, and we now prove that the reverse inclusion also holds. By Lemma 11.74, each element $u \in \mathcal{O}_E$ has an expression

$$u = c_0 + c_1\zeta + c_2\zeta^2 + \cdots + c_{p-2}\zeta^{p-2},$$

where $c_i \in \mathbb{Q}$ (remember that $[E:\mathbb{Q}] = p-1$). We must show that $c_i \in \mathbb{Z}$ for all i. Multiplying by $1-\zeta$ gives

$$u(1-\zeta) = c_0(1-\zeta) + c_1(\zeta - \zeta^2) + \cdots + c_{p-2}(\zeta^{p-2} - \zeta^{p-1}).$$

By (i), $\text{tr}(\zeta^i - \zeta^{i+1}) = \text{tr}(\zeta^i) - \text{tr}(\zeta^{i+1}) = 0$ for $1 \le i \le p-2$, so that $\text{tr}(u(1-\zeta)) = c_0\,\text{tr}(1-\zeta)$; hence, $\text{tr}(u(1-\zeta) = pc_0$, because $\text{tr}(1-\zeta) = p$, by (ii). On the other hand, $\text{tr}(u(1-\zeta)) \in p\mathbb{Z}$, by (iv). Hence, $pc_0 = mp$ for some $m \in \mathbb{Z}$, and so $c_0 \in \mathbb{Z}$. Now $\zeta^{-1} = \zeta^{p-1} \in \mathcal{O}_E$, so that

$$(u - c_0)\zeta^{-1} = c_1 + c_2\zeta + \cdots + c_{p-2}\zeta^{p-3} \in \mathcal{O}_E.$$

The argument just given shows that $c_1 \in \mathbb{Z}$. Indeed, repetition of this argument shows that all $c_i \in \mathbb{Z}$, and so $u \in \mathbb{Z}[\zeta]$. •

Before we leave this interesting topic, we must mention a beautiful theorem of Dirichlet. For proofs of the following statements, see Samuel, *Algebraic Theory of Numbers*, Chapter 4. An algebraic number field E of degree n has exactly n imbeddings into $\mathbb{C}$. If r_1 is the number of such imbeddings with image in $\mathbb{R}$, then $n - r_1$ is even; say, $n - r_1 = 2r_2$.

Theorem (Dirichlet Unit Theorem). *Let E be an algebraic number field of degree n. Then $n = r_1 + 2r_2$, (where r_1 is the number of imbeddings of E into $\mathbb{R}$), and the multiplicative group $U(\mathcal{O}_E)$ of units in $\mathcal{O}_E$ is a finitely generated abelian group. More precisely,*

$$U(\mathcal{O}_E) \cong \mathbb{Z}^{r_1+r_2-1} \times T,$$

where T is a finite cyclic group consisting of the roots of unity in E.

EXERCISES

11.41 (i) If $E = \mathbb{Q}(\sqrt{-3})$, prove that the only units in $\mathcal{O}_E$ are

$$\pm 1, \quad \tfrac{1}{2}(1 \pm \sqrt{-3}), \quad \tfrac{1}{2}(-1 \pm \sqrt{-3}),$$

(ii) Let d be a negative squarefree integer with $d \neq -1$ and $d \neq -3$. If $E = \mathbb{Q}(\sqrt{d})$, prove that the only units in $\mathcal{O}_E$ are ± 1.

11.42 (i) Prove that if $E = \mathbb{Q}(\sqrt{2}) \subseteq \mathbb{R}$, then there are no units $u \in \mathcal{O}_E$ with $1 < u < 1 + \sqrt{2}$.

(ii) If $E = \mathbb{Q}(\sqrt{2})$, prove that $\mathcal{O}_E$ has infinitely many units.

Hint. Use (i) to prove that all powers of $1 + \sqrt{2}$ are distinct.

Definition. If $\mathcal{O}_E$ is the ring of integers in an algebraic number field E, then a ***discriminant*** of $\mathcal{O}_E$ is

$$\Delta(\mathcal{O}_E) = \prod_{i<j} (\alpha_i - \alpha_j)^2,$$

where $\alpha_1, \ldots, \alpha_n$ is an integral basis of $\mathcal{O}_E$.

11.43 Let d be a squarefree integer, and let $E = \mathbb{Q}(\sqrt{d})$.

(i) If $d \equiv 2 \bmod 4$ or $d \equiv 3 \bmod 4$, prove that $1, \sqrt{d}$ is an integral basis of $\mathcal{O}_E$, and prove that a discriminant of $\mathcal{O}_E$ is $4d$.

(ii) If $d \equiv 1 \bmod 4$, prove that $1, \frac{1}{2}(1 + \sqrt{d})$ is an integral basis of $\mathcal{O}_E$, and prove that a discriminant of $\mathcal{O}_E$ is d.

11.44 Let p be an odd prime, and let $E = \mathbb{Q}(\zeta_p)$ be the cyclotomic field.

(i) Show that $1, 1 - \zeta_p, (1 - \zeta_p)^2, \ldots, (1 - \zeta_p)^{p-2}$ is an integral basis for $\mathcal{O}_E$.

(ii) Prove that a discriminant of $\mathcal{O}_E$ is $(-1)^{\frac{1}{2}(p-1)} p^{p-2}$.

Hint. See Pollard, *The Theory of Algebraic Numbers*, page 67.

11.45 (i) If $\mathbb{A}$ is the field of all algebraic numbers, prove that $\mathcal{O}_{\mathbb{A}}$ is not noetherian.

(ii) Prove that every nonzero prime ideal in $\mathcal{O}_{\mathbb{A}}$ is a maximal ideal.

Hint. Use the proof of Corollary 11.53.

Characterizations of Dedekind Rings

The following definition involves some of the ring-theoretic properties enjoyed by the ring of integers $\mathcal{O}_E$ in an algebraic number field E.

Definition. A domain R is a ***Dedekind ring*** if it is integrally closed, noetherian, and its nonzero prime ideals are maximal ideals.

Example 11.88.

(i) The ring $\mathcal{O}_E$ in an algebraic number field E is a Dedekind ring, by Proposition 11.46, Proposition 11.84, and Corollary11.53.

(ii) Every principal ideal domain R is a Dedekind ring. ◂

It is shown, in Example 11.85, that $R = \mathbb{Z}[\sqrt{-5}\,]$ is a Dedekind ring that is not a UFD and, hence, it is not a PID. We remind the reader that E. Kummer, in his investigations into Fermat's last theorem in the 1840s, recognized such examples, and he forced unique factorization by adjoining "ideal" numbers to rings of integers. About 30 years later, R. Dedekind introduced the modern definition of ideal, and showed that Kummer's ideal numbers correspond to Dedekind's ideals. We will prove, in Theorem 11.95, that every nonzero ideal in a Dedekind ring has a unique factorization as a product of prime ideals.

We now characterize DVRs, and then show that localizations of Dedekind rings are well-behaved.

Lemma 11.89. *A domain R is a* DVR *if and only if it is noetherian, integrally closed, and has a unique nonzero prime ideal.*

Proof. If R is a DVR, then it does have the required properties (recall that R is a PID, hence it is integrally closed).

The converse, which requires us to show that R is a PID, is not as simple as we would expect. Let $\mathfrak{p}$ be the nonzero prime ideal, and choose a nonzero $a \in \mathfrak{p}$. Define $M = R/Ra$, and consider the family $\mathcal{A}$ of all the annihilators $\operatorname{ann}(m)$ as m varies over all the nonzero elements of M. Since R is noetherian, it satisfies the maximum condition, and so there is a nonzero element $b + Ra \in M$ whose annihilator $\mathfrak{q} = \operatorname{ann}(b + Ra)$ is maximal in $\mathcal{A}$. We claim that $\mathfrak{q}$ is a prime ideal. Suppose that $x, y \in R$, $xy \in \mathfrak{q}$, and $x, y \notin \mathfrak{q}$. Then $y(b + Ra) = yb + Ra$ is a nonzero element of M, because $y \notin \mathfrak{q}$. But $\operatorname{ann}(yb + Ra) \supsetneq \operatorname{ann}(b + Ra)$, because $x \notin \operatorname{ann}(b + Ra)$, contradicting the maximality property of $\mathfrak{q}$. Therefore, $\mathfrak{q}$ is a prime ideal. Since R has a unique nonzero prime ideal $\mathfrak{p}$, we have

$$\mathfrak{q} = \operatorname{ann}(b + Ra) = \mathfrak{p}.$$

Note that

$$b/a \notin R.$$

Otherwise, $b+Ra = 0+Ra$, contradicting $b+Ra$ being a nonzero element of $M = R/Ra$.

We now show that $\mathfrak{p}$ is principal, with generator a/b (we do not yet know whether $a/b \in \operatorname{Frac}(R)$ lies in R). First, we have $\mathfrak{p}b = \mathfrak{q}b \subseteq Ra$, so that $\mathfrak{p}(b/a) \subseteq R$; that is, $\mathfrak{p}(b/a)$ is an ideal in R. If $\mathfrak{p}(b/a) \subseteq \mathfrak{p}$, then b/a is integral over R, for $\mathfrak{p}$ is a finitely generated R-submodule of $\operatorname{Frac}(R)$, as required in Lemma 11.41. As R is integrally closed, this puts $b/a \in R$, contradicting what we noted at the end of the previous paragraph. Therefore, $\mathfrak{p}(b/a)$ is not a proper ideal, so that $\mathfrak{p}(b/a) = R$ and $\mathfrak{p} = R(a/b)$. It follows that $a/b \in R$ and $\mathfrak{p}$ is a principal ideal.

Denote a/b by t. The proof is completed by showing that the only nonzero ideals in R are the principal ideals generated by t^n, for $n \geq 0$. Let I be a nonzero ideal in R, and

consider the chain of submodules of $\operatorname{Frac}(R)$:

$$I \subseteq It^{-1} \subseteq It^{-2} \subseteq \cdots .$$

We claim that this chain is strictly increasing. If $It^{-n} = It^{-n-1}$, then the finitely generated R-module It^{-1} satisfies $t^{-1}(It^{-n}) \subseteq It^{-n}$, so that $t^{-1} = b/a$ is integral over R. As above, R integrally closed forces $b/a \in R$, a contradiction. Since R is noetherian, this chain can contain only finitely many ideals in R. Thus, there is n with $It^{-n} \subseteq R$ and $It^{-n-1} \not\subseteq R$. If $It^{-n} \subseteq \mathfrak{p} = Rt$, then $It^{-n-1} \subseteq R$, a contradiction. Therefore, $It^{-n} = R$ and $I = Rt^n$, as desired. •

Proposition 11.90. *If R is a noetherian domain, then R is a Dedekind ring if and only if for every nonzero prime ideal $\mathfrak{p}$, the localization $R_\mathfrak{p}$ is a* DVR.

Remark. Exercise 11.45 on page 948 shows that it is necessary to assume that R is noetherian. ◂

Proof. If R is a Dedekind ring and $\mathfrak{p}$ is a maximal ideal, Corollary 11.18(iv) shows that $R_\mathfrak{p}$ has a unique nonzero prime ideal. Moreover, $R_\mathfrak{p}$ is noetherian (Corollary 11.18(v)), a domain (Corollary 11.16), and integrally closed (Exercise 11.26). By Lemma 11.89, $R_\mathfrak{p}$ is a DVR.

For the converse, we must show that R is integrally closed and that its nonzero prime ideals are maximal. Let $u/v \in \operatorname{Frac}(R)$ be integral over R. For every nonzero prime ideal $\mathfrak{p}$, the element u/v is integral over $R_\mathfrak{p}$ (note that $\operatorname{Frac}(R_\mathfrak{p}) = \operatorname{Frac}(R)$). But $R_\mathfrak{p}$ is a PID, hence is integrally closed, and so $u/v \in R_\mathfrak{p}$. We conclude that $u/v \in \bigcap_\mathfrak{p} R_\mathfrak{p} = R$, by Proposition 11.20. Therefore, R is integrally closed.

Suppose there were nonzero prime ideals $\mathfrak{p} \subsetneq \mathfrak{q}$ in R. By Corollary 11.18(iv), $\mathfrak{p}_\mathfrak{q} \subsetneq \mathfrak{q}_\mathfrak{q}$ in $R_\mathfrak{q}$. This contradicts the fact that a DVR has a unique nonzero prime ideal. Therefore, nonzero prime ideals are maximal, and so R is a Dedekind ring. •

Let R be a domain with $F = \operatorname{Frac}(R)$, and let $I = Ra$ be a nonzero principal ideal in R. If we define $J = Ra^{-1} \subseteq F$, the cyclic R-submodule generated by a^{-1}, then it easy to see that

$$IJ = \{uv : u \in I \text{ and } v \in J\} = R.$$

Definition. If R is a domain with $F = \operatorname{Frac}(R)$, then a ***fractional ideal*** is a finitely generated nonzero R-submodule of F. If I is a nonzero ideal in R, then

$$I^{-1} = \{v \in F : vI \subseteq R\}.$$

It is always true that $I^{-1}I \subseteq R$; a fractional ideal I is ***invertible*** if $I^{-1}I = R$.

Every finitely generated ideal in R is also a fractional ideal. In this context, we often call such ideals (which are the usual ideals!) ***integral ideals*** when we want to contrast them with more general fractional ideals.

We claim that if $I = Ra$ is a nonzero principal ideal in R, then $I^{-1} = Ra^{-1}$. Clearly, $(ra^{-1})(r'a) = rr' \in R$ for all $r' \in R$, so that $Ra^{-1} \subseteq I^{-1}$. For the reverse inclusion, suppose that $(u/v)a \in R$, where $u, v \in R$. Then $v \mid ua$ in R, so there is $r \in R$ with $rv = ua$. Hence, in F, we have $u = rva^{-1}$, so that $u/v = (rva^{-1})/v = ra^{-1}$. Therefore, every nonzero principal ideal in R is invertible.

Lemma 11.91. *If R is a domain with $F = \mathrm{Frac}(R)$, then a fractional ideal I is invertible if and only if there exist $a_1, \ldots, a_n \in I$ and $q_1, \ldots, q_n \in F$ with*

(i) $q_i I \subseteq R$ *for* $i = 1, \ldots, n$;

(ii) $1 = \sum_{i=1}^n q_i a_i$.

Proof. If I is invertible, then $I^{-1}I = R$. Since $1 \in I^{-1}I$, there are $a_1, \ldots, a_n \in R$ and $q_1, \ldots, q_n \in I^{-1}$ with $1 = \sum_i q_i a_i$. Since $q_i \in I^{-1}$, we have $q_i I \subseteq R$.

To prove the converse, the R-submodule J of F generated by $q_1, \ldots, q_n$ is a fractional ideal. Since $1 = \sum_{i=1}^n q_i a_i \in JI$, JI is an R-submodule of R containing 1; that is, $JI = R$. To see that I is invertible, it remains to prove that $J = I^{-1}$. Clearly, each $q_i \in I^{-1}$, so that $J \subseteq I^{-1}$. For the reverse inclusion, assume that $u \in F$ and $uI \subseteq R$. Since $1 = \sum_i q_i a_i$, we have $u = \sum_i (ua_i) q_i \in J$ because $ua_i \in R$ for all i. •

Corollary 11.92. *Every invertible ideal I in a domain R is finitely generated.*

Proof. Since I is invertible, there exist $a_1, \ldots, a_n \in I$ and $q_1, \ldots, q_n \in F$ as in the lemma. If $b \in I$, then $b = b1 = \sum_i bq_i a_i \in I$, because $bq_i \in R$. Therefore, I is generated by $a_1, \ldots, a_n \in I$. •

Proposition 11.93. *The following conditions are equivalent for a domain R.*

(i) *R is a Dedekind ring.*

(ii) *Every fractional ideal is invertible.*

(iii) *The set of all the fractional ideals $\mathcal{F}(R)$ forms an abelian group under multiplication of ideals.*

Proof. (i) ⇒ (ii).

Let J be a fractional ideal in R. Since R is a Dedekind ring, its localization $R_\mathfrak{p}$ is a PID, and so $J_\mathfrak{p}$, as every nonzero principal ideal, is invertible (in Theorem 9.3, in the course of proving that finitely generated torsion-free abelian groups are free abelian, we really proved that fractional ideals of PIDs are cyclic modules). Now Exercise 11.50 on page 958 gives

$$(J^{-1}J)_\mathfrak{p} = (J^{-1})_\mathfrak{p} J_\mathfrak{p} = (J_\mathfrak{p})^{-1} J_\mathfrak{p} = R_\mathfrak{p}.$$

Proposition 11.30 gives $J^{-1}J = R$, and so J is invertible.

(ii) ⇔ (iii).

If $I, J \in \mathcal{F}(R)$, then they are finitely generated, by Corollary 11.92, and

$$IJ = \left\{\sum a_\ell b_\ell : a_\ell \in I \text{ and } b_\ell \in J\right\}$$

is a finitely generated R-submodule of $\text{Frac}(R)$. If $I = (a_1, \ldots, a_n)$ and $J = (b_1, \ldots, b_m)$, then IJ is generated by all $a_i b_j$. Hence, IJ is finitely generated and $IJ \in \mathcal{F}(R)$. Associativity does hold, the identity is R, and the inverse of a fractional ideal J is J^{-1}, because J is invertible. It follows that $\mathcal{F}(R)$ is an abelian group.

Conversely, if $\mathcal{F}(R)$ is an abelian group and $I \in \mathcal{F}(R)$, then there is $J \in \mathcal{F}(R)$ with $JI = R$. We must show that $J = I^{-1}$. But

$$R = JI \subseteq I^{-1}I \subseteq R,$$

so that $JI = I^{-1}I$. Canceling I in the group $\mathcal{F}(R)$ gives $J = I^{-1}$, as desired.

(iii) $\Rightarrow$ (i).

First, R is noetherian, for (iii) $\Rightarrow$ (ii) shows that every nonzero ideal I is invertible, and Corollary 11.92 shows that I is finitely generated.

Second, we show that every nonzero prime ideal $\mathfrak{p}$ is a maximal ideal. Let I be an ideal with $\mathfrak{p} \subsetneq I$ (we allow $I = R$). Then $\mathfrak{p}I^{-1} \subseteq II^{-1} = R$, so that $\mathfrak{p}I^{-1}$ is an (integral) ideal in R. Now $(\mathfrak{p}I^{-1})I = \mathfrak{p}$, because multiplication is associative in $\mathcal{F}(R)$. Since $\mathfrak{p}$ is a prime ideal, Proposition 6.13 says that either $\mathfrak{p}I^{-1} \subseteq \mathfrak{p}$ or $I \subseteq \mathfrak{p}$. The second option does not hold, so that $\mathfrak{p}I^{-1} \subseteq \mathfrak{p}$. Multiplying by $\mathfrak{p}^{-1}I$ gives $R \subseteq I$. Therefore, $I = R$, and so $\mathfrak{p}$ is a maximal ideal.

Third, if $a \in \text{Frac}(R)$ is integral over R, then Lemma 11.41 gives a finitely generated R-submodule J of $\text{Frac}(R)$, i.e., a fractional ideal, with $aJ \subseteq J$. Since J is invertible, there are $q_1, \ldots, q_n \in \text{Frac}(R)$ and $a_1, \ldots, a_n \in J$ with $q_i J \subseteq R$ for all i and $1 = \sum q_i a_i$. Hence, $a = \sum_i q_i a_i a$. But $a_i a \in J$ and $q_i J \subseteq R$ gives $a = \sum_i q_i(a_i a) \in R$. Therefore, R is integrally closed, and hence it is a Dedekind ring. •

Proposition 11.94.

(i) *If R is a* UFD, *then a nonzero ideal I in R is invertible if and only if it is principal.*

(ii) *A Dedekind ring R is a* UFD *if and only if it is a* PID.

Proof. (i) We have already seen that every nonzero principal ideal is invertible. Conversely, if I is invertible, there are elements $a_1, \ldots, a_n \in I$ and $q_1, \ldots, q_n \in \text{Frac}(R)$ with $1 = \sum_i q_i a_i$ and $q_i I \subseteq R$ for all i. Let $q_i = b_i/c_i$, where $b_, c_i \in R$. Since R is a UFD, we may assume that q_i is in lowest terms; that is, $(b_i, c_i) = 1$. But $(b_i/c_i)a_j \in R$ says that $c_i \mid b_i a_j$, so that $c_i \mid a_j$ for all i, j, by Exercise 6.18(i) on page 339. We claim that $I = Rc$, where $c = \text{lcm}\{c_1, \ldots, c_n\}$. First, $c \in I$, for $cb_i/c_i \in R$ and $c = c1 = \sum_i (cb_i/c_i)a_i$. Hence, $Rc \subseteq I$. For the reverse inclusion, Exercise 6.18(ii) on page 339 shows that $c \mid a_j$ for all j, so that $a_j \in Rc$, for all j, and so $I \subseteq Rc$.

(ii) Since every nonzero ideal in a Dedekind ring is invertible, it follows from (i) that if R is a UFD, then every ideal in R is principal. •

Definition. If R is a Dedekind ring, then its ***class group*** $C(R)$ is defined by

$$C(R) = \mathcal{F}(R)/\mathcal{P}(R),$$

where $\mathcal{P}(R)$ is the subgroup of all nonzero principal ideals.

Dirichlet proved, for every algebraic number field E, that the class group of $C(\mathcal{O}_E)$ is finite; the order $|C(R)|$ is called the ***class number*** of $\mathcal{O}_E$. The usual proof of finiteness of the class number uses a geometric theorem of H. Minkowski which says that sufficiently large parallelopipeds in euclidean space must contain lattice points (see Samuel, *Algebraic Theory of Numbers*, pages 57–58).

L. Claborn proved, for every (not necessarily finite) abelian group G, that there is a Dedekind ring R with $C(R) \cong G$.

We can now prove the result linking Kummer and Dedekind.

Theorem 11.95. *If R is a Dedekind ring, then every proper nonzero ideal has a unique factorization as a product of prime ideals.*

Proof. Let $\mathcal{S}$ be the family of all proper nonzero ideals in R that are not products of prime ideals. If $\mathcal{S} = \varnothing$, then every nonzero ideal in R is a product of prime ideals. If $\mathcal{S} \neq \varnothing$, then $\mathcal{S}$ has a maximal element I, because noetherian rings satisfy the maximum condition (Proposition 6.38). Now I cannot be a maximal ideal in R, for a "product of prime ideals" is allowed to have only one factor. Let $\mathfrak{m}$ be a maximal ideal containing I. Since $I \subsetneq \mathfrak{m}$, we have $\mathfrak{m}^{-1}I \subsetneq \mathfrak{m}^{-1}\mathfrak{m} = R$; that is, $\mathfrak{m}^{-1}I$ is a proper ideal properly containing I. Neither $\mathfrak{m}$ nor $\mathfrak{m}^{-1}I$ lies in $\mathcal{S}$, for each is strictly larger than a maximal element, namely, I, and so each of them is a product of prime ideals. Therefore, $I = \mathfrak{m}(\mathfrak{m}^{-1}I)$ (equality holding because R is a Dedekind ring) is a product of prime ideals, contradicting I being in $\mathcal{S}$. Therefore, $\mathcal{S} = \varnothing$, and every proper nonzero ideal in R is a product of prime ideals.

Suppose that $\mathfrak{p}_1 \cdots \mathfrak{p}_r = \mathfrak{q}_1 \cdots \mathfrak{q}_s$, where the $\mathfrak{p}_i$ and $\mathfrak{q}_j$ are prime ideals. We prove unique factorization by induction on $\max\{r, s\}$. The base step $r = 1 = s$ is obviously true. For the inductive step, note that $\mathfrak{p}_1 \supseteq \mathfrak{q}_1 \cdots \mathfrak{q}_s$, so that Proposition 6.13 gives $\mathfrak{q}_j$ with $\mathfrak{p}_1 \supseteq \mathfrak{q}_j$. Hence, $\mathfrak{p}_1 = \mathfrak{q}_j$, because prime ideals are maximal. Now multiply the original equation by $\mathfrak{p}_1^{-1}$ and use the inductive hypothesis. •

Corollary 11.96. *If R is a Dedekind ring, then $\mathcal{F}(R)$ is a free abelian group with basis all the nonzero prime ideals.*

Proof. Of course, $\mathcal{F}(R)$ is written multiplicatively. That every fractional ideal is a product of primes shows that the set of primes generates $\mathcal{F}(R)$; uniqueness of the factorization says the set of primes is a basis. •

In light of Theorem 11.95, many of the usual formulas of arithmetic extend to ideals in Dedekind rings. Observe that in $\mathbb{Z}$, the ideal (3) contains (9). In fact, $\mathbb{Z}m \supseteq \mathbb{Z}n$ if and only if $m \mid n$. We will now see that the relation "contains" for ideals is the same as "divides," and that the usual formulas for gcd's and lcm's (Proposition 1.17) generalize to Dedekind rings.

Proposition 11.97. *Let I and J be nonzero ideals in a Dedekind ring R, and let their prime factorizations be*

$$I = \mathfrak{p}_1^{e_1} \cdots \mathfrak{p}_n^{e_n} \qquad and \qquad J = \mathfrak{p}_1^{f_1} \cdots \mathfrak{p}_n^{f_n},$$

where $e_i \geq 0$ and $f_i \geq 0$ for all i.

(i) *$J \supseteq I$ if and only if $I = JL$ for some ideal L.*

(ii) *$J \supseteq I$ if and only if $f_i \leq e_i$ for all i.*

(iii) *If $m_i = \min\{e_i, f_i\}$ and $M_i = \max\{e_i, f_i\}$, then*

$$I \cap J = \mathfrak{p}_1^{M_1} \cdots \mathfrak{p}_n^{M_n} \qquad and \qquad I + J = \mathfrak{p}_1^{m_1} \cdots \mathfrak{p}_n^{m_n}.$$

In particular, $I + J = R$ if and only if $\min\{e_i, f_i\} = 0$ for all i.

(iv) *Let R be a Dedekind ring, and let $I = \mathfrak{p}_1^{e_1} \cdots \mathfrak{p}_n^{e_n}$ be a nonzero ideal in R. Then*

$$R/I = R/\mathfrak{p}_1^{e_1} \cdots \mathfrak{p}_n^{e_n} \cong \big(R/\mathfrak{p}_1^{e_1}\big) \times \cdots \times \big(R/\mathfrak{p}_n^{e_n}\big).$$

Proof. (i) If $I \subseteq J$, then $J^{-1}I \subseteq R$, and

$$J(J^{-1}I) = I.$$

Conversely, if $I = JL$, then $I \subseteq J$ because $JL \subseteq JR = J$.

(ii) This follows from (i) and the unique factorization of nonzero ideals as products of prime ideals.

(iii) We prove the formula for $I + J$. Let $I + J = \mathfrak{p}_1^{r_1} \cdots \mathfrak{p}_n^{r_n}$ and let $A = \mathfrak{p}_1^{m_1} \cdots \mathfrak{p}_n^{m_n}$. Since $I \subseteq I + J$ and $J \subseteq I + J$, we have $r_i \leq e_i$ and $r_i \leq f_i$, so that $r_i \leq \min\{e_i, f_i\} = m_i$. Hence, $A \subseteq I + J$. For the reverse inclusion, $A \subseteq I$ and $A \subseteq J$, so that $A = II'$ and $A = JJ'$ for ideals I' and J', by (i). Therefore, $I + J = AI' + AJ' = A(I' + J')$, and so $I + J \subseteq A$. The proof of the formula for IJ is left to the reader.

(iv) This is just the Chinese remainder theorem, Exercise 6.11(iii) on page 325, so that it suffices to verify the hypothesis that $\mathfrak{p}_i^{e_i}$ and $\mathfrak{p}_j^{e_j}$ are coprime when $i \neq j$; that is, $\mathfrak{p}_i^{e_i} + \mathfrak{p}_j^{e_j} = R$. But this follows from (iii). •

Recall Proposition 7.58: an R-module A is projective if and only if it has a *projective basis*: there exist elements $\{a_j : j \in J\} \subseteq A$ and R-maps $\{\varphi_j : A \to R : j \in J\}$ such that

(i) for each $x \in A$, almost all $\varphi_j(x) = 0$;

(ii) for each $x \in A$, we have $x = \sum_{j \in J} (\varphi_j x) a_j$.

Proposition 11.98.

(i) *A nonzero ideal I in a domain R is invertible if and only if I is a projective R-module.*

(ii) *A domain R is a Dedekind ring if and only if every ideal in R is projective.*

Proof. (i) If I is invertible, there are elements $a_1, \ldots, a_n \in I$ and $q_1, \ldots, q_n \in \text{Frac}(R)$ with $1 = \sum_i q_i a_i$ and $q_i I \subset R$ for all i. Define $\varphi_i : I \to R$ by $\varphi_i : a \mapsto q_i a$ (note that $\text{im}\, \varphi_i \subseteq I$ because $q_i I \subseteq R$). If $a \in I$, then

$$\sum_i \varphi_i(a) a_i = \sum_i q_i a a_i = a \sum_i q_i a_i = a.$$

Therefore, I has a projective basis, and so I is a projective R-module.

Conversely, if I is a projective, it has a projective basis $\{\varphi_j : j \in J\}, \{a_j : j \in J\}$. If $b \in I$ is nonzero, define $q_j \in \text{Frac}(R)$ by

$$q_j = \varphi_j(b)/b.$$

This element does not depend on the choice of nonzero b: if $b' \in I$ is nonzero, then $b'\varphi_j(b) = \varphi_j(b'b) = b\varphi_j(b')$, so that $\varphi_j(b)/b = \varphi_j(b')/b'$. To see that $q_j I \subseteq R$, note that if $b \in I$ is nonzero, then $q_j b = \big(\varphi_j(b)/b\big)b = \varphi_j(b) \in R$. By item (i) in the definition of projective basis, almost all $\varphi_j(b) = 0$, and so there are only finitely many nonzero $q_j = \varphi_j(b)/b$ (remember that q_j does not depend on the choice of nonzero $b \in I$). Item (ii) in the definition of projective basis gives, for $b \in I$,

$$b = \sum_j \varphi_j(b) a_j = \sum_j (q_j b) a_j = b\big(\sum_j q_j a_j\big).$$

Canceling b gives $1 = \sum_j q_j a_j$. Finally, the set of those a_j with indices j for which $q_j \neq 0$ completes the data necessary to show that I is an invertible ideal.

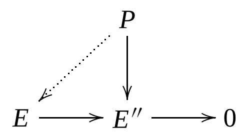
(ii) This follows at once from (i) and Proposition 11.93. •

Example 11.99.
We have seen that $R = \mathbb{Z}\big[\sqrt{-5}\,\big]$ is a Dedekind ring that is not a PID. Any non-principal ideal gives an example of a projective R-module that is not free. ◂

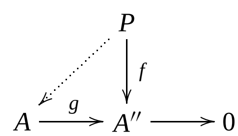
Remark. A not necessarily commutative ring R is called ***left hereditary*** if every left ideal is a projective R-module (there exist rings that are left hereditary but not right hereditary). Some examples of left hereditary rings aside from Dedekind rings are semisimple rings, noncommutative principal ideal rings, and FIRs (*free ideal rings*—all left ideals are free R-modules). P. M. Cohn proved that polynomial rings over a field in noncommuting variables are FIRs, and so there exist left hereditary rings that are not left noetherian. ◂

The projective and injective modules over a Dedekind ring are well-behaved.

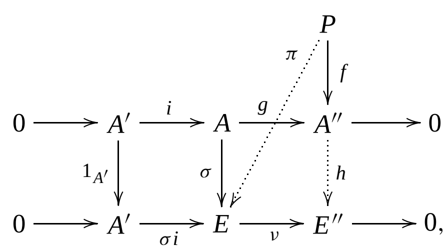
Lemma 11.100. *A left R-module P (over any ring R) is projective if and only if every diagram below with E injective can be completed to a commutative diagram. The dual characterization of injective modules is also true.*



Proof. If P is projective, then the diagram can be completed for every not necessarily injective module E. Conversely, we must show that the diagram



can be completed for any module A and any surjection $g: A \to A''$. By Theorem 8.104, there is an injective R-module E and an injection $\sigma: A \to E$. Define $E'' = \operatorname{coker} \sigma i = E/\operatorname{im} \sigma i$, and consider the commutative diagram with exact rows



where $\nu: E \to E'' = \operatorname{coker} \sigma i$ is the natural map and $h: A'' \to E''$ exists by Proposition 8.93. By hypothesis, there exists a map $\pi: P \to E$ with $\nu\pi = hf$. We claim that $\operatorname{im} \pi \subseteq \operatorname{im} \sigma$. For $x \in P$, surjectivity of g gives $a \in A$ with $ga = fx$. Then $\nu\pi x = hfx = hga = \nu\sigma a$, and so $\pi x - \sigma a \in \ker \nu = \operatorname{im} \sigma i$; hence, $\pi x - \sigma a = \sigma i a'$ for some $a' \in A'$, and so $\pi x = \sigma(a + ia') \in \operatorname{im} \sigma$. Therefore, if $x \in P$, there is a unique $a \in A$ with $\sigma a = \pi x$ (a is unique because σ is an injection). Thus, there is a well-defined function $\pi': P \to A$, given by $\pi' x = a$, where $\sigma a = \pi x$. The reader may check that π' is an R-map and that $g\pi' = f$. •

Theorem 11.101 (Cartan-Eilenberg). *The following conditions are equivalent for a domain R.*

(i) *R is a Dedekind ring.*

(ii) *Every submodule of a projective R-module is projective.*

(iii) *Every quotient of an injective R-module is injective.*

Proof. (i) $\Leftrightarrow$ (ii).

If R is Dedekind, then we can adapt the proof of Theorem 9.8 (which proves that every subgroup of a free abelian group is free abelian) to prove that every submodule of a free R-module is projective (see Exercise 11.47 on page 958); in particular, every submodule of a projective R-module is projective. Conversely, since R itself is a projective R-module, its submodules are also projective, by hypothesis; that is, the ideals of R are projective. Proposition 11.98 now shows that R is a Dedekind ring.

(ii) $\Leftrightarrow$ (iii).

Assume (iii), and consider the diagram with exact rows

$$\begin{array}{ccccc} P & \longleftarrow & P' & \longleftarrow & 0 \\ \downarrow & \searrow & \downarrow f & & \\ E & \longrightarrow & E'' & \longrightarrow & 0, \end{array}$$

where P is projective and E is injective; note that the hypothesis gives E'' injective. To prove projectivity of P', it suffices, by Lemma 11.100, to find a map $P' \to E$ making the diagram commute. Since E'' is injective, there exists a map $P \to E''$ giving commutativity. Since P is projective, there is a map $P \to E$ also giving commutativity. The composite $P' \to P \to E$ is the desired map. The converse is the dual of this, using the dual of Lemma 11.100. •

Corollary 11.102. *Let R be a Dedekind ring.*

(i) *For all R-modules C and A, then for all $n \geq 2$,*

$$\mathrm{Ext}_R^n(C, A) = \{0\} \qquad \textit{and} \qquad \mathrm{Tor}_n^R(C, A) = \{0\}.$$

(ii) *Let $0 \to A' \to A \to A'' \to 0$ be a short exact sequence. For every module C, there are exact sequences*

$$0 \to \mathrm{Hom}(C, A') \to \mathrm{Hom}(C, A) \to \mathrm{Hom}(C, A'')$$
$$\to \mathrm{Ext}^1(C, A') \to \mathrm{Ext}^1(C, A) \to \mathrm{Ext}^1(C, A'') \to 0$$

and

$$0 \to \mathrm{Tor}_1^R(C, A') \to \mathrm{Tor}_i^R(C, A) \to \mathrm{Tor}_1^R(C, A'')$$
$$\to C \otimes_R A' \to C \otimes_R A \to C \otimes_R A'' \to 0$$

Proof. (i) By definition, if

$$\cdots \to P_2 \xrightarrow{d_2} P_1 \xrightarrow{d_1} P_0 \to C \to 0$$

is a projective resolution of C, then

$$\mathrm{Ext}^n(C, A) = \ker d_{n+1}^* / \operatorname{im} d_n^*;$$

moreover, $\text{Ext}^n(C, A)$ does not depend on the choice of projective resolution, by Corollary 10.74. Now C is a quotient of a free module F, and so there is an exact sequence

$$0 \to K \to F \xrightarrow{\varepsilon} C \to 0, \tag{5}$$

where $K = \ker \varepsilon$. Since R is Dedekind, the submodule K of the free R-module F is projective, so that (5) defines a projective resolution of C with $P_0 = F$, $P_1 = K$, and $P_n = \{0\}$ for all $n \geq 2$. Hence, $\ker d^*_{n+1} \subseteq \text{Hom}(P_n, A) = \{0\}$ for all $n \geq 2$, and so $\text{Ext}^n_R(C, A) = \{0\}$ for all A and for all $n \geq 2$. A similar argument works for Tor.

(ii) This follows from Corollary 10.68 and Corollary 10.57, the long exact sequence for Ext and for Tor, respectively. •

We will use this result in the next section to generalize Proposition 10.92.

EXERCISES

11.46 Let R be a commutative ring and let M be a finitely generated R-module. Prove that if $IM = M$ for some ideal I of R, then there exists $a \in J$ with $(1 - a)M = \{0\}$.
Hint. If $M = \langle m_1 \dots, m_n \rangle$, then each $m_i = \sum_j a_{ij} m_j$, where $a_{ij} \in I$. Use the adjoint matrix (the matrix of cofactors) as in the proof of Lemma 11.41.

11.47 Generalize the proof of Theorem 9.8 to prove that if R is a left hereditary ring, then every submodule of a free left R-module F is isomorphic to a direct sum of ideals, and hence is projective.

11.48 Let R be a Dedekind ring, and let $\mathfrak{p}$ be a nonzero prime ideal in R.
(i) If $a \in \mathfrak{p}$, prove that $\mathfrak{p}$ occurs in the prime factorization of Ra.
(ii) If $a \in \mathfrak{p}^e$ and $a \notin \mathfrak{p}^{e+1}$, prove that $\mathfrak{p}^e$ occurs in the prime factorization of Ra, but that $\mathfrak{p}^{e+1}$ does not occur in the prime factorization of Ra.

11.49 Let I be a nonzero ideal in a Dedekind ring R. Prove that if $\mathfrak{p}$ is a prime ideal, then $I \subseteq \mathfrak{p}$ if and only if $\mathfrak{p}$ occurs in the prime factorization of I.

11.50 If J is a fractional ideal of a Dedekind ring R, prove that $(J^{-1})_\mathfrak{p} = (J_\mathfrak{p})^{-1}$ for every maximal ideal $\mathfrak{p}$.

11.51 Let $I_1, \dots, I_n$ be ideals in a Dedekind ring R. If there is no nonzero prime ideal $\mathfrak{p}$ with $I_i = \mathfrak{p}L_i$ for all i for ideals L_i, then

$$I_1 + \cdots + I_n = R.$$

11.52 Give an example of a projective $\mathbb{Z}[\sqrt{-5}]$-module that is not free.
Hint. See Example 11.99.

11.53 (i) A commutative ring R is called a ***principal ideal ring*** if every ideal in R is a principal ideal (R would be a PID if it were a domain). For example, $\mathbb{I}_n$ is a principal ideal ring. Prove that $\mathbb{Z} \times \mathbb{Z}$ is not a principal ideal ring.
(ii) Let $I_1, \dots, I_n$ be pairwise coprime ideals in a commutative ring R. If R/I_i is a principal ideal ring for each i, prove that $R/(I_1 \cdots I_n)$ is a principal ideal ring.
Hint. Use the Chinese remainder theorem, Exercise 6.11(iii) on page 325.

11.54 Let a be a nonzero element in a Dedekind ring R. Prove that there are only finitely many ideals I in R containing a.
Hint. If $a \in I$, then $Ra = IL$ for some ideal $L \subseteq R$.

Finitely Generated Modules over Dedekind Rings

We saw, in Chapter 9, that theorems about abelian groups generalize to theorems about modules over PIDs. We are now going to see that such theorems can be further generalized to modules over Dedekind rings.

Proposition 11.103. *Let R be a Dedekind ring.*

(i) *If $I \subseteq R$ is a nonzero ideal, then every ideal in R/I is principal.*

(ii) *Every fractional ideal J can be generated by two elements. More precisely, for any nonzero $a \in J$, there exists $b \in J$ with $J = Ra + Rb$.*

Proof. (i) Let $I = \mathfrak{p}_1^{e_1} \cdots \mathfrak{p}_n^{e_n}$ be the prime factorization of I. Since the ideals $\mathfrak{p}_i^{e_i}$ are pairwise coprime, it suffices, by Exercise 11.53(ii) on page 958, to prove that $R/\mathfrak{p}_i^{e_i}$ is a principal ideal ring for each i. Now right exactness of $R_{\mathfrak{p}_i} \otimes_R$ shows that $(R/\mathfrak{p}_i^{e_i})_{\mathfrak{p}_i} \cong R_{\mathfrak{p}_i}/(\mathfrak{p}_i^{e_i})_{\mathfrak{p}_i}$. Since $R_{\mathfrak{p}_i}$ is a PID (it is even a DVR), any quotient ring of it is a principal ideal ring.

(ii) Assume first that J is an integral ideal. Choose a nonzero $a \in J$. By (i), the ideal J/Ra in R/Ra is principal; say, J/Ra is generated by $b + Ra$, where $b \in J$. It follows that $J = Ra + Rb$.

For the general case, there is a nonzero $c \in R$ with $cJ \subseteq R$ (if J is generated by $u_1/v_1, \ldots, u_m/v_m$, take $c = \prod_J v_j$). Since cJ is an integral ideal, given any nonzero $a \in J$, there is $cb \in cJ$ with $cJ = Rca + Rcb$. It follows that $J = Ra + Rb$. •

The next corollary says that we can force nonzero ideals to be coprime.

Corollary 11.104. *If I and J are fractional ideals over a Dedekind ring R, then there are $a, b \in \mathrm{Frac}(R)$ with*

$$aI + bJ = R.$$

Proof. Choose a nonzero $a \in I^{-1}$. Now $aI \subseteq I^{-1}I = R$, so that $aIJ^{-1} \subseteq J^{-1}$. By Proposition 11.103(ii), there is $b \in J^{-1}$ with

$$J^{-1} = aIJ^{-1} + Rb.$$

Since $b \in J^{-1}$, we have $bJ \subseteq R$, and so

$$R = JJ^{-1} = J(aIJ^{-1} + Rb) = aI + RbJ = aI + bJ. \quad \bullet$$

Let us now investigate the structure of R-modules.

Lemma 11.105.

(i) *If* $0 \to M' \to M \to M'' \to 0$ *is a short exact sequence of* R*-modules, then*

$$\operatorname{rank}(M) = \operatorname{rank}(M') + \operatorname{rank}(M'').$$

(ii) *An* R*-module is torsion if and only if* $\operatorname{rank}(M) = 0$.

(iii) *If* M *is a finitely generated torsion-free* R*-module with* $M \neq \{0\}$, *then* $\operatorname{rank}(M) = 1$ *if and only if* M *is isomorphic to a nonzero ideal.*

Proof. (i) By Corollary 8.103, the fraction field F is a flat R-module. Therefore, $0 \to F \otimes_R M' \to F \otimes_R M \to F \otimes_R M'' \to 0$ is a short exact sequence of vector spaces over F, and the result is a standard result of linear algebra (Exercise 3.74 on page 171).

(ii) If M is torsion, then $F \otimes_R M = \{0\}$, by an obvious generalization of Proposition 8.95 (divisible $\otimes$ torsion $= \{0\}$). Hence, $\operatorname{rank}(M) = 0$. Conversely, if $\operatorname{rank}(M) = 0$, then $F \otimes_R M = \{0\}$. By Proposition 11.25, if $S = R - \{0\}$ and $h_M : M \to S^{-1}M$ is the localization map, then $\ker h_M = \{m \in M : \sigma m = 0 \text{ for some } \sigma \in R - \{0\}\}$. Thus, $M = \ker h_M$ here, and so M is torsion.

(iii) If $M \cong I$, where I is an ideal, then $\operatorname{rank}(M) = \operatorname{rank}(I)$, and there is an exact sequence $0 \to I \to F$. Since F is a flat R-module, the sequence $0 \to F \otimes_R I \to F \otimes_R F$ is exact. But $F \otimes_R F \cong F$ is one-dimensional, so that $\operatorname{rank}(I) \leq 1$. As $I \neq \{0\}$ (because fractional ideals are nonzero), we have $\operatorname{rank}(I) = 1$.

Conversely, assume that $\operatorname{rank}(M) = 1$; that is, $F \otimes_R M \cong F$. Choose nonzero elements $u, v \in M$. If u, v are linearly independent, then $\langle u, v \rangle = \langle u \rangle \oplus \langle v \rangle$. But exactness of $0 \to \langle u \rangle \oplus \langle v \rangle \to M$ gives exactness of $0 \to F \otimes_R \langle u \rangle \oplus F \otimes_R \langle v \rangle \to F \otimes_R M$ (we have used the flatness of F once again). This is a contradiction, for a one-dimensional space has no two-dimensional subspaces. Choose a nonzero element $x \in M$. It follows that if $m \in M$, then there exist nonzero $r, s \in R$ with $sm = rx$. The reader may adapt the argument in the proof of Theorem 9.3 to see that the function $M \to F$, given by $m \mapsto r/s$, is a well-defined (because M is torsion-free) isomorphism of M and a submodule S of F. As M is finitely generated, S is a fractional ideal.

It remains to show that every fractional ideal $J = \langle a_1/b_n, \ldots, a_n/b_n \rangle \subseteq F$ is isomorphic to an integral ideal. If $b = \prod_i b_i$, then $bx \in R$ for all $x \in J$, for multiplication by b merely clears denominators. Hence, the map $J \to R$, given by $x \mapsto bx$, is an R-map; it is injective because fields have no zero divisors. •

Proposition 11.106. *If* R *is a Dedekind ring and* M *is a finitely generated torsion-free* R*-module, then*

$$M \cong I_1 \oplus \cdots \oplus I_n,$$

where I_i *is an ideal in* R.

Proof. The proof is by induction on $\operatorname{rank}(M) \geq 0$. If $\operatorname{rank}(M) = 0$, then M is torsion, by Lemma 11.105(ii). Since M is torsion-free, $M = \{0\}$. Assume now that $\operatorname{rank}(M) = n + 1$.

Choose a nonzero $m \in M$, so that $\text{rank}(Rm) = 1$. The sequence

$$0 \to Rm \to M \xrightarrow{\nu} M'' \to 0$$

is exact, where $M'' = R/Rm$ and ν is the natural map. Note that $\text{rank}(M'') = n$, by Lemma 11.105(i). Now M finitely generated implies M'' is also finitely generated. If $T = t(M'')$ is the torsion submodule of M'', then M''/T is a finitely generated torsion-free R-module with $\text{rank}(M''/T) = \text{rank}(M'') = n$, because $\text{rank}(T) = 0$. By induction, M''/T is a direct sum of ideals, hence is projective. Define

$$M' = \nu^{-1}(T) = \{m \in M : rm \in Rm \text{ for some } r \neq 0\} \subseteq M.$$

There is an exact sequence $0 \to M' \to M \to M''/T \to 0$; this sequence splits because M''/T is projective; that is, $M \cong M' \oplus (M''/T)$. Hence

$$\text{rank}(M') = \text{rank}(M) - \text{rank}(M''/T) = 1.$$

Since R is noetherian, every submodule of a finitely generated R-module is itself finitely generated; hence, M' is finitely generated. Therefore, M' is isomorphic to an ideal, by Lemma 11.105(ii), and this completes the proof. •

Corollary 11.107. *If R is a Dedekind ring and M is a finitely generated torsion-free R-module, then M is projective.*

Proof. Recall that every ideal in a Dedekind ring is projective, by Proposition 11.98. It now follows from Proposition 11.106 that M is a direct sum of ideals, and hence it is projective.

We can also prove this result using localization. For every maximal ideal $\mathfrak{m}$, the $R_\mathfrak{m}$-module $M_\mathfrak{m}$ is finitely generated torsion-free. Since $R_\mathfrak{m}$ is a PID (even a DVR), however, $M_\mathfrak{m}$ is a free module, and hence it is projective. The result now follows from Corollary 11.39. •

Corollary 11.108. *If M is a finitely generated R-module, where R is a Dedekind ring, then the torsion submodule tM is a direct summand of M.*

Proof. The quotient module M/tM is a finitely generated torsion-free R-module, so that it is projective, by Corollary 11.107. Therefore, tM is a direct summand of M, by Corollary 7.55. •

Corollary 11.109. *If R is a Dedekind ring, then every torsion-free R-module A is flat.*

Proof. By Lemma 8.97, it suffices to prove that every finitely generated submodule of A is flat. But such submodules are torsion-free, hence projective, and projective modules are always flat, by Lemma 8.98. •

It can be proved, over an arbitrary domain R, that every flat R-module is torsion-free (see Rotman, *An Introduction to Homological Algebra*, page 129).

Using homological algebra, we generalize Corollary 11.108 by removing the hypothesis that tM be finitely generated.

Corollary 11.110. *Let R be a Dedekind ring with $F = \mathrm{Frac}(R)$.*

(i) *If C is a torsion-free R-module and T is a torsion module with $\mathrm{ann}(T) \neq \{0\}$, then $\mathrm{Ext}^1_R(C, T) = \{0\}$.*

(ii) *Let M be an R-module. If $\mathrm{ann}(tM) \neq \{0\}$, where tM is the torsion submodule of M, then tM is a direct summand of M.*

Proof. We generalize the proof of Proposition 10.92. Since C is torsion-free, it is a flat R-module, by Corollary 11.109, so that exactness of $0 \to R \to F$ gives exactness of $0 \to R \otimes_R C \to F \otimes_R C$. Thus, $C \cong R \otimes_R C$ can be imbedded in a vector space V over F, namely, $V = F \otimes_R C$. Applying the contravariant functor $\mathrm{Hom}_R(\ , T)$ to $0 \to C \to V \to V/C \to 0$ gives an exact sequence

$$\mathrm{Ext}^1_R(V, T) \to \mathrm{Ext}^1_R(C, T) \to \mathrm{Ext}^2(V/C, T).$$

Now the last term is $\{0\}$, by Corollary 11.102, and $\mathrm{Ext}^1_R(V, T)$ is (torsion-free) divisible, by (a straightforward generalization of) Example 10.70, so that $\mathrm{Ext}^1_R(C, T)$ is divisible. Since $\mathrm{ann}(T) \neq \{0\}$, Exercise 10.41 on page 852 gives $\mathrm{Ext}^1_R(C, T) = \{0\}$.

(i) To prove that the extension $0 \to tM \to M \to M/tM \to 0$ splits, it suffices to prove that $\mathrm{Ext}^1_R(M/tM, tM) = \{0\}$. Since M/tM is torsion-free, this follows from part (i) and Corollary 10.90. •

The next result generalizes Proposition 7.73.

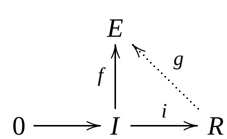
Proposition 11.111. *The following statements are equivalent for a domain R.*

(i) *R is a Dedekind ring.*

(ii) *An R-module E is injective if and only if it is divisible.*

Proof. (i) $\Rightarrow$ (ii).

Let R be a Dedekind ring and let E be a divisible R-module. By the Baer criterion, Theorem 7.68, it suffices to complete the diagram



where I is an ideal and $i : I \to R$ is the inclusion. Of course, we may assume that I is nonzero, and so I is invertible: there are elements $a_1, \ldots, a_n \in I$ and elements $q_1 \ldots, q_n \in F$ with $q_i I \subseteq R$ and $1 = \sum_i q_i a_i$. Since E is divisible, there are elements $e_i \in E$ with $f(a_i) = a_i e_i$. Note, for every $b \in I$, that

$$f(b) = f\Big(\sum_i q_i a_i b\Big) = \sum_i (q_i b)\, f(a_i) = \sum_i (q_i b) a_i e_i = b \sum_i (q_i a_i) e_i.$$

Hence, if we define $e = \sum_i (q_i a_i)e_i$, then $e \in E$ and $f(b) = be$ for all $b \in I$. Defining $g\colon R \to E$ by $g(r) = re$ shows that the diagram can be completed, and so E is injective. That every injective R-module is divisible was proved (for arbitrary domains R) in Lemma 7.72.

(ii) ⇒ (i).

Let E be an injective R-module. If E' is a quotient of E, then E' is divisible and hence, by hypothesis, injective. Therefore, every quotient of an injective module is injective, and so R is a Dedekind ring, by Theorem 11.101. •

Having examined torsion-free modules, let us now look at torsion modules.

Proposition 11.112. *Let $\mathfrak{p}$ be a nonzero prime ideal in a Dedekind ring R. If M is an R-module with* $\operatorname{ann}(M) = \mathfrak{p}^e$ *for some $e > 0$, then the localization map $M \to M_\mathfrak{p}$ is an isomorphism (and hence M may be regarded as an $R_\mathfrak{p}$-module).*

Proof. It suffices to prove that $M \cong R_\mathfrak{p} \otimes_R M$. If $m \in M$ is nonzero and $s \in R$ with $s \notin \mathfrak{p}$, then

$$\mathfrak{p}^e + Rs = R,$$

by Proposition 11.97. Hence, there exist $u \in \mathfrak{p}^e$ and $r \in R$ with $1 = u + rs$, and so

$$m = um + rsm = rsm.$$

If $1 = u' + r's$, where $u' \in \mathfrak{p}$ and $r' \in R$, then $s(r - r')m = 0$, so that

$$s(r - r') \in \operatorname{ann}(m) = \mathfrak{p}^e.$$

Since $s \notin \mathfrak{p}^e$, it follows that $r - r' \in \mathfrak{p}^e$ (the prime factorization of Rs does not contain $\mathfrak{p}^e$; if the prime factorization of $R(r - r')$ does not contain $\mathfrak{p}^e$, then neither does the prime factorization of $Rs(r-r')$). Hence, $rm = r'm$. Define $s^{-1}m = rm$. Define $f\colon R_\mathfrak{p} \times M \to M$ by $f(r/s, m) = s^{-1}rm$, where $s^{-1}rm$ has been defined in the preceding paragraph. It is straightforward to check that f is R-bilinear, and so there is an R-map $\widetilde{f}\colon R_\mathfrak{p} \otimes M \to M$ with $\widetilde{f}\colon (r/s \otimes m) = s^{-1}rm$. In particular, $\widetilde{f}(1 \otimes m) = m$, so that $\widetilde{f}$ is surjective. On the other hand, the localization map $h_M\colon M \to R_\mathfrak{p} \otimes_R M$, defined by $h_M(m) = 1 \otimes m$, is easily seen to be the inverse of $\widetilde{f}$. •

Definition. Let $\mathfrak{p}$ be a nonzero prime ideal in a Dedekind ring R. An R-module M is called $\mathfrak{p}$-***primary*** if, for each $m \in M$, there is $e > 0$ with $\operatorname{ann}(m) = \mathfrak{p}^e$.

Theorem 11.113 (Primary Decomposition). *Let R be a Dedekind ring, and let T be a finitely generated torsion R-module. If $I = \operatorname{ann}(T) = \mathfrak{p}_1^{e_1} \cdots \mathfrak{p}_n^{e_n}$, then*

$$T = T[\mathfrak{p}_1] \oplus \cdots \oplus T[\mathfrak{p}_n],$$

where

$$T[\mathfrak{p}_i] = \{m \in M : \operatorname{ann}(m) \text{ is a power of } \mathfrak{p}_i\}.$$

*$T[\mathfrak{p}_i]$ is called the $\mathfrak{p}_i$-**primary component** of T.*

Proof. It is easy to see that the $\mathfrak{p}_i$-primary components $T[\mathfrak{p}_i]$ are submodules of T. We now check the conditions in Proposition 7.19. If W_i is the submodule of T generated by all $T[\mathfrak{p}_j]$ with $j \neq i$, we must show that $T[\mathfrak{p}_i] \cap W_i = \{0\}$. Let $x \in T[\mathfrak{p}_i] \cap W_i$. If

$$I_i = \mathfrak{p}_1^{e_1} \cdots \widehat{\mathfrak{p}_i^{e_i}} \cdots \mathfrak{p}_n^{e_n},$$

then $\mathfrak{p}_i$ and I_i are coprime: $\mathfrak{p}_i + I_i = R$. Hence, there are $a_i \in \mathfrak{p}_i$ and $r_i \in I_i$ with $1 = a_i + r_i$, and so $x = a_i x + r_i x$. But $a_i x = 0$, because $x \in T[\mathfrak{p}_i]$, and $r_i x = 0$, because $x \in W_i$ and $I_i = \operatorname{ann}(W_i)$. Therefore, $x = 0$.

By Exercise 11.51 on page 958, we have

$$I_1 + \cdots + I_n = R.$$

Thus, there are $b_i \in I_i$ with $b_1 + \cdots + b_n = 1$. If $t \in T$, then $t = b_1 t + \cdots + b_n t$. But if $c_i \in \mathfrak{p}_i^{e_i}$, then $c_i b_i \in \mathfrak{p}_i^{e_i} I_i = I = \operatorname{ann}(T)$ and so $c_i(b_i t) = 0$. Hence, $\mathfrak{p}_i^{e_i} \subseteq \operatorname{ann}(b_i t)$, so that $\operatorname{ann}(b_i t) = \mathfrak{p}_i^e$ for some $e > 0$. Therefore, $b_i t \in T[\mathfrak{p}_i]$, and so

$$T = T[\mathfrak{p}_1] + \cdots + T[\mathfrak{p}_n].$$

The result now follows from Proposition 7.19. •

Theorem 11.114. *Let R be a Dedekind ring.*

(i) *Two finitely generated torsion R-modules T and T' are isomorphic if and only if $T[\mathfrak{p}_i] \cong T'[\mathfrak{p}_i]$ for all i.*

(ii) *Every finitely generated $\mathfrak{p}$-primary R-module T is a direct sum of cyclic R-modules, and the number of summands of each type is an invariant of T.*

Proof. (i) The result follows easily from the observation that if $f : T \to T'$ is an isomorphism, then $\operatorname{ann}(t) = \operatorname{ann}(f(t))$ for all $t \in T$.

(ii) The primary decomposition shows that T is the direct sum of its primary components $T[\mathfrak{p}_i]$. By Proposition 11.112, $T[\mathfrak{p}_i]$ is an $R_{\mathfrak{p}_i}$-module. But $R_{\mathfrak{p}_i}$ is a PID (even a DVR), and so the basis theorem and the fundamental theorem hold: each $T[\mathfrak{p}_i]$ is a direct sum of cyclic modules, and the numbers and isomorphism types of the cyclic summands are uniquely determined. •

We now know that every finitely generated R-module M is a direct sum of cyclic modules and ideals. What uniqueness is there in such a decomposition? Since the torsion submodule is a fully invariant direct summand, we may focus on torsion-free modules.

Recall Proposition 11.3: Two ideals J and J' in a domain R are isomorphic if and only if there is $a \in \operatorname{Frac}(R)$ with $J' = aJ$.

Lemma 11.115. *Let M be a finitely generated torsion-free R-module, where R is a Dedekind ring, so that $M \cong I_1 \oplus \cdots \oplus I_n$, where the I_i are ideals. Then*

$$M \cong R^{n-1} \oplus J,$$

where $J = I_1 \cdots I_n$.

Remark. We call $R^{n-1} \oplus J$ a ***Steinitz normal form*** for M. We will prove, in Theorem 11.117, that J is unique up to isomorphism. ◂

Proof. It suffices to prove that $I \oplus J \cong R \oplus IJ$, for the result then follows easily by induction on $n \geq 2$. By Corollary 11.104, there are nonzero $a, b \in \mathrm{Frac}(R)$ with $aI+bJ = R$. Since $aI \cong I$ and $bJ \cong J$, we may assume that I and J are coprime integral ideals. There is an exact sequence

$$0 \to I \cap J \xrightarrow{\delta} I \oplus J \xrightarrow{\alpha} I + J \to 0,$$

where $\delta \colon x \mapsto (x, x)$ and $\alpha \colon (u, v) \mapsto u - v$. Since I and J are coprime, however, we have $I \cap J = IJ$ and $I + J = R$. As R is projective, this sequence splits; that is, $I \oplus J \cong R \oplus IJ$. •

The following cancellation lemma, while true for Dedekind rings, can be false for some other rings. In Example 7.78(iii), we described an example of Swan showing that if $R = \mathbb{R}[x_1, \ldots, x_n]/(1-\sum_i x_i^2)$ is the real coordinate ring of the 3-sphere, then there is a finitely generated stably free R-module M that is not free. Hence, there are free R-modules F and F' with $M \oplus F \cong F' \oplus F$ but $M \ncong F'$.

Lemma 11.116. *Let R be a Dedekind ring. If $R \oplus G \cong R \oplus H$, where G and H are R-modules, then $G \cong H$.*

Proof. We may assume there is a module $E = A \oplus G = B \oplus H$, where $A \cong R \cong B$. Let $p \colon E \to B$ be the projection $p \colon (b, h) \mapsto b$, and let $p' = p|G$. Now

$$\ker p' = G \cap H \qquad \text{and} \qquad \mathrm{im}\, p' \subseteq B \cong R.$$

Thus, $\mathrm{im}\, p' \cong L$, where L is an ideal in R.

If $\mathrm{im}\, p' = \{0\}$, then $G \subseteq \ker p = H$. Since $E = A \oplus G$, Corollary 7.18 gives $H = G \oplus (H \cap A)$. On the one hand, $E/G = (A \oplus G)/G \cong A \cong R$; on the other hand, $E/G = (B \oplus H)/G \cong B \oplus (H/G) \cong R \oplus (H/G)$. Thus, $R \cong R \oplus (H/G)$. Since R is a domain, this forces $H/G = \{0\}$: if $R = X \oplus Y$, then X and Y are ideals; if $x \in X$ and $y \in Y$ are both nonzero, then $xy \in X \cap Y = \{0\}$, giving zero divisors in R. It follows that $H/G = \{0\}$ and $G = H$.

We may now assume that $L = \mathrm{im}\, p'$ is a nonzero ideal. The first isomorphism theorem gives $G/(G \cap H) \cong L$. Since R is a Dedekind ring, L is a projective module, and so

$$G = I \oplus (G \cap H),$$

where $I \cong L$. Similarly,

$$H = J \oplus (G \cap H),$$

where J is isomorphic to an ideal. Therefore,

$$E = A \oplus G = A \oplus I \oplus (G \cap H);$$
$$E = B \oplus H = B \oplus J \oplus (G \cap H).$$

It follows that

$$A \oplus I \cong E/(G \cap H) \cong B \oplus J.$$

If we can prove that $I \cong J$, then

$$G = I \oplus (G \cap H) \cong J \oplus (G \cap H) = H.$$

Therefore, we have reduced the theorem to the special case when G and H are nonzero ideals.

We will prove that if $\alpha\colon R \oplus I \to R \oplus J$ is an isomorphism, then $I \cong J$. As in our discussion of generalized matrices on page 540, α determines a 2×2 matrix

$$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix},$$

where $a_{11}\colon R \to R$, $a_{21}\colon R \to J$, $a_{12}\colon I \to R$, and $a_{22}\colon I \to J$. Indeed, as maps between ideals are just multiplications by elements of $F = \mathrm{Frac}(R)$, we may regard A as a matrix in $\mathrm{GL}(2, F)$. Now $a_{21} \in J$ and $a_{22}I \subseteq J$, so that if $d = \det(A)$, then

$$dI = (a_{11}a_{22} - a_{12}a_{12})I \subseteq J.$$

Similarly, $\beta = \alpha^{-1}$ determines a 2×2 matrix $B = A^{-1}$.

$$d^{-1}J = \det(B)J \subseteq I,$$

so that $J \subseteq dI$. We conclude that $J = dI$, and so $J \cong I$. •

Let us sketch a proof using exterior algebra that if R is a Dedekind ring and I and J are fractional ideals, then $R \oplus I \cong R \oplus J$ implies $I \cong J$. The fact that 2×2 determinants are used in the original proof suggests that second exterior powers may be useful. By Theorem 9.143,

$$\bigwedge^2(R \oplus I) \cong \Big(R \otimes \bigwedge^2(I)\Big) \oplus \Big(\bigwedge^1(R) \otimes_R \bigwedge^1(I)\Big) \oplus \Big(\bigwedge^2(R) \otimes_R I\Big).$$

Now $\bigwedge^2(R) = \{0\}$, by Corollary 9.138, and $\bigwedge^1(R) \otimes_R \bigwedge^1(I) \cong R \otimes_R I \cong I$. We now show, for every maximal ideal $\mathfrak{m}$, that $\big(\bigwedge^2(I)\big)_{\mathfrak{m}} = \{0\}$. By Exercise 11.24 on page 921,

$$\Big(\bigwedge^n(I)\Big)_{\mathfrak{m}} \cong \bigwedge^n(I_{\mathfrak{m}}).$$

But $R_{\mathfrak{m}}$ is a PID, so that $I_{\mathfrak{m}}$ is a principal ideal, and hence $\bigwedge^2(I_{\mathfrak{m}}) = \{0\}$, by Corollary 9.138. It now follows from Proposition 11.31(i) that $\bigwedge^2(I) = \{0\}$. Therefore, $\bigwedge^2(R \oplus I) \cong I$. Similarly, $\bigwedge^2(R \oplus J) \cong J$, and so $I \cong J$.

Theorem 11.117 (Steinitz). *Let R be a Dedekind ring, and let $M \cong I_1 \oplus \cdots I_n$ and $M' \cong I'_1 \oplus \cdots I'_\ell$ be finitely generated torsion-free R-modules. Then $M \cong M'$ if and only if $n = \ell$ and $I_1 \cdots I_n \cong I'_1 \cdots I'_n$.*

Proof. Lemma 11.105(iii) shows that $\text{rank}(I_i) = 1$ for all i, and Lemma 11.105(i) shows that $\text{rank}(M) = n$; similarly, $\text{rank}(M') = \ell$. Since $M \cong M'$, we have $F \otimes_R M \cong F \otimes_R M'$, so that $\text{rank}(M) = \text{rank}(M')$ and $n = \ell$. By Lemma 11.115, it suffices to prove that if $R^n \oplus I \cong R^n \oplus J$, then $I \cong J$. But this follows at once from repeated use of Lemma 11.116. •

Let R be a commutative ring, and let $\mathcal{C}$ be a subcategory of ${}_R\textbf{Mod}$. Recall that two R-modules A and B are called *stably isomorphic in* $\mathcal{C}$ if there exists a module $C \in \text{obj}(\mathcal{C})$ with $A \oplus C \cong B \oplus C$.

Corollary 11.118. *Let R be a Dedekind ring, and let $\mathcal{C}$ be the category of all finitely generated torsion-free R-modules. Then $M, M' \in \mathcal{C}$ are stably isomorphic in $\mathcal{C}$ if and only if they are isomorphic.*

Proof. Isomorphic modules are always stably isomorphic. To prove the converse, assume that there is a finitely generated torsion-free R-module X with

$$M \oplus X \cong M' \oplus X.$$

There are ideals I, J, L with $M \cong R^{n-1} \oplus I$, $M' \cong R^{n-1} \oplus J$, and $X \cong R^{m-1} \oplus L$, where $n = \text{rank}(M) = \text{rank}(M')$. Hence,

$$M \oplus X \cong R^{n-1} \oplus I \oplus R^{m-1} \oplus L \cong R^{n+m-1} \oplus IL.$$

Similarly,

$$M' \oplus X \cong R^{n+m-1} \oplus JL.$$

By Theorem 11.117, $IL \cong JL$, and so there is a nonzero $a \in \text{Frac}(R)$ with $aIL = JL$, by Lemma 11.3. Multiplying by L^{-1} gives $aI = J$, and so $I \cong J$. Therefore,

$$M \cong R^{n-1} \oplus I \cong R^{n-1} \oplus J \cong M'. \quad \bullet$$

Recall that if a category $\mathcal{C}$ has finite products, then the *Grothendieck group* $K_0(\mathcal{C})$ is the abelian group with generators (the isomorphism classes of) $\text{obj}(\mathcal{C})$ and relations $A \oplus B = A + B$ for all $A, B \in \text{obj}(\mathcal{C})$; that is, $K_0(\mathcal{C}) = \mathcal{F}(\mathcal{C})/\mathcal{R}$, where $\mathcal{F}(\mathcal{C})$ is the free abelian group with basis $\text{obj}(\mathcal{C})$ and $\mathcal{R}$ is the subgroup generated by all $A \oplus B - A - B$. If $[A]$ denotes the element $A + \mathcal{R}$ in $K_0(\mathcal{C})$, where $A \in \text{obj}(\mathcal{C})$, then $[A] = [B]$ in $K_0(\mathcal{C})$ if and only if they are stably isomorphic in $\mathcal{C}$, by Proposition 7.77.

Notation. If $\textbf{Pr}(R)$ is the category of all finitely generated projective R-modules over a commutative ring R, write

$$K_0(R) = K_0(\textbf{Pr}(R)).$$

We end this section by displaying a relation between the class group $C(R)$ of a Dedekind ring R and its Grothendieck group $K_0(R)$. If I is a nonzero ideal in a Dedekind ring R, denote the corresponding element in $C(R)$ by $\text{cls}(I)$.

Theorem 11.119. *If R is a Dedekind ring, then*

$$K_0(R) \cong C(R) \oplus \mathbb{Z},$$

where $C(R)$ is the class group of R.

Proof. If P is a finitely generated projective R-module, then $P \cong R^{n-1} \oplus I$ for a nonzero ideal I, by Lemma 11.115; moreover, the isomorphism class of I is uniquely determined by P, by Theorem 11.117. If $P \cong R^{n-1} \oplus J$, then there is $a \in \mathrm{Frac}(R)$ with $J = aI$, so that $\mathrm{cls}(J) = \mathrm{cls}(I)$ in $C(R)$. Therefore, the function $\varphi\colon K_0(R) \to C(R) \oplus \mathbb{Z}$, given by

$$\varphi([P]) = (\mathrm{cls}(I), \mathrm{rank}(P)),$$

is well-defined. Note that we are writing the first summand $C(R)$ multiplicatively and the second summand $\mathbb{Z}$ additively. To see that φ is well-defined on $K_0(R) = \mathcal{F}(\mathbf{Pr}(R))/\mathcal{R}$, it suffices to prove that it preserves the relations in $\mathcal{R}$; that is,

$$\varphi([P \oplus Q]) - \varphi([P]) - \varphi([Q]) = 0.$$

Let $Q = R^{m-1} \oplus J$, where J is a nonzero ideal. Then $P \oplus Q \cong R^{n+m-1} \oplus IJ$, and

$$\begin{aligned}\varphi([P \oplus Q]) &= (\mathrm{cls}(IJ), n+m)\\ &= (\mathrm{cls}(I)\mathrm{cls}(J), n+m)\\ &= (\mathrm{cls}(I), n) + (\mathrm{cls}(J), m).\end{aligned}$$

Since $\mathrm{rank}(P \oplus Q) = \mathrm{rank}(P) + \mathrm{rank}(Q)$, it follows that φ is a well-defined homomorphism.

Now φ is surjective, for $(\mathrm{cls}(I), n) = \varphi([R^{n-1} \oplus I])$, and $C(R) \oplus \mathbb{Z}$ is generated by all such elements. To see that φ is injective, recall that Proposition 7.77 says that a typical element of $K_0(R)$ has the form $[P] - [Q]$. If $\varphi([P] - [Q]) = 0$, then $\varphi([P]) = \varphi([Q])$. Hence, Proposition 7.77 says that P and Q are stably isomorphic. Corollary 11.118 says that $P \cong Q$, and so $[P] - [Q] = 0$. Therefore, φ is an isomorphism. •

Remark. There is another Grothendieck group in this context. An R-module A is called ***invertible*** if it is finitely generated and $A \otimes_R \mathrm{Hom}_R(A, R) \cong R$. By Propositions 8.83 and 9.97, the category of all invertible R-modules under tensor product is a $\star$-category, and so it has a Grothendieck group, which is called the ***Picard group***, denoted by $\mathrm{Pic}(R)$. It turns out that every invertible module is isomorphic to an ideal. Thus, $\mathrm{Pic}(R)$ is the abelian group (written multiplicately) with generators all invertible ideals in R and relations $I \otimes_R J = IJ$. When R is a Dedekind ring, then $\mathrm{Pic}(R) \cong C(R)$. ◂

EXERCISES

11.55 Let R be a Dedekind ring, and let $I \subseteq R$ be a nonzero ideal. Prove that there exists an ideal $J \subseteq R$ with $I + J = R$ and IJ principal.

Hint. Let $I = \mathfrak{p}_1^{e_1} \cdots \mathfrak{p}_n^{e_n}$, and choose $r_i \in \mathfrak{p}_i^{e_i} - \mathfrak{p}_i^{e_i+1}$. Use the Chinese remainder theorem to find an element $a \in R$ with $a \in \mathfrak{p}_i^{e_i}$ and $a \notin \mathfrak{p}_i^{e_i+1}$, and consider the prime factorization of Ra.

11.56 (i) If R is a commutative ring, prove that $R^n \cong R^m$ implies $n = m$.

Hint. If $\mathfrak{m}$ is a maximal ideal in R, then the $(R/\mathfrak{m})$-vector spaces $(R/\mathfrak{m})^n$ and $(R/\mathfrak{m})^m$ are isomorphic.

(ii) If R is any commutative ring, prove that $\mathbb{Z}$ is a direct summand of $K_0(R)$.

11.57 If R is a PID, prove that $K_0(R) \cong \mathbb{Z}$.

11.58 If I is a fractional ideal in a Dedekind ring R, prove that $I \otimes I^{-1} \cong R$.
Hint. Use invertibility of I.

11.59 If R is a local ring, prove that $K_0(R) \cong \mathbb{Z}$.

11.60 If R is a commutative ring, prove that rank defines a surjective homomorphism $K_0(R) \to \mathbb{Z}$. We usually call the kernel of this map the ***reduced*** Grothendieck group, and we denote it by $\widetilde{K}_0(R)$. Hence,

$$K_0(R) \cong \widetilde{K}_0(R) \oplus \mathbb{Z}.$$

11.61 If $\mathcal{C}$ is a subcategory of ${}_R\mathbf{Mod}$, then we defined a variant of the Grothendieck group on page 492: $K'(\mathcal{C})$ is the abelian group with generators obj$(\mathcal{C})$ and relations $B = A - C$ if there exists a (not necessarily split) exact sequence $0 \to A \to B \to C \to 0$.

(i) If R is a Dedekind ring, prove that restriction of the homomorphism $\varepsilon \colon K_0(R) \to K'(\mathcal{C})$ of Proposition 7.82 is an isomorphism $\widetilde{K}_0(R) \to K'(\mathcal{C})$.

(ii) If R is a Dedekind ring, prove that $K'(\mathcal{C}) \cong C(R)$.

11.3 Global Dimension

There are several types of rings whose finitely generated modules have been classified: semisimple rings; PIDs; Dedekind rings. Each of these rings can be characterized in terms of its projective modules: a ring R is semisimple if and only if every R-module is projective; a domain R is Dedekind if and only if every ideal is projective. The notion of global dimension allows us to classify arbitrary rings in this spirit.

Rings in this section need not be commutative.

Definition. Let R be a ring and let A be a left R-module. If there is a finite projective resolution

$$0 \to P_n \to \cdots \to P_1 \to P_0 \to A \to 0,$$

then we write $pd(A) \leq n$. If $n \geq 0$ is the smallest integer such that $pd(A) \leq n$, then we say that A has ***projective dimension*** n; if there is no finite projective resolution of A, then $pd(A) = \infty$.

Example 11.120.
(i) A module A is projective if and only if $pd(A) = 0$. We may thus regard $pd(A)$ as a measure of how far away A is from being projective.

(ii) If R is a Dedekind ring, then $pd(A) \leq 1$ for every R-module A. By Theorem 11.101, every submodule of a free R-module is projective. Hence, if F is a free R-module and $\varepsilon \colon F \to A$ is a surjection, then

$$0 \to \ker \varepsilon \to F \xrightarrow{\varepsilon} A \to 0$$

is a projective resolution of A. This argument extends to left hereditary rings. ◂

Definition. Let $\mathbf{P}_\bullet = \cdots \to P_2 \xrightarrow{d_2} P_1 \xrightarrow{d_1} P_0 \xrightarrow{\varepsilon} A \to 0$ be a projective resolution of a module A. If $n \geq 0$, then the nth ***syzygy*** is

$$\Omega_n(A, \mathbf{P}_\bullet) = \begin{cases} \ker \varepsilon & \text{if } n = 0 \\ \ker d_n & \text{if } n \geq 1. \end{cases}$$

Proposition 11.121. *For every $n \geq 1$, for all left R-modules A and B, and for every projective resolution $\mathbf{P}_\bullet$ of B, there is an isomorphism*

$$\text{Ext}_R^{n+1}(A, B) \cong \text{Ext}_R^1(\Omega_{n-1}(A, \mathbf{P}_\bullet), B).$$

Proof. The proof is by induction on $n \geq 1$. Exactness of the projective resolution

$$\mathbf{P}_\bullet = \cdots \to P_2 \xrightarrow{d_2} P_1 \xrightarrow{d_1} P_0 \xrightarrow{\varepsilon} A \to 0$$

gives exactness of

$$\cdots \to P_2 \xrightarrow{d_2} P_1 \to \Omega_0(A, \mathbf{P}_\bullet) \to 0,$$

which is a projective resolution $\mathbf{P}_\bullet^+$ of $\Omega_0(A, \mathbf{P}_\bullet)$. In more detail, define

$$P_n^+ = P_{n+1} \qquad \text{and} \qquad d_n^+ = d_{n+1}.$$

Since Ext^1 is independent of the choice of projective resolution of the first variable,

$$\text{Ext}_R^1(\Omega_0(A, \mathbf{P}_\bullet), B) = \frac{\ker(d_2^+)^*}{\text{im}(d_1^+)^*} = \frac{\ker(d_3)^*}{\text{im}(d_2)^*} = \text{Ext}_R^2(A, B).$$

The inductive step is proved in the same way, noting that

$$\cdots \to P_{n+2} \xrightarrow{d_{n+2}} P_{n+1} \to \Omega_n(A, \mathbf{P}_\bullet) \to 0$$

is a projective resolution of $\Omega_n(A, \mathbf{P}_\bullet)$. •

Corollary 11.122. *For all left R-modules A and B, for all $n \geq 0$, and for any projective resolutions $\mathbf{P}_\bullet$ and $\mathbf{P}'_\bullet$ of A, there is an isomorphism*

$$\mathrm{Ext}^1_R(\Omega_n(A, \mathbf{P}_\bullet), B) \cong \mathrm{Ext}^1_R(\Omega_n(A, \mathbf{P}'_\bullet), B).$$

Proof. By Proposition 11.121. both are isomorphic to $\mathrm{Ext}^{n+1}_R(A, B)$. •

Two modules Ω and Ω' are called ***projectively equivalent*** if there exist projective modules P and P' with $\Omega \oplus P \cong \Omega' \oplus P'$. Exercise 11.62 on page 983 shows that any two nth syzygies of a module A are projectively equivalent. We often abuse notation and speak of *the* nth syzygy of a module, writing $\Omega_n(A)$ instead of $\Omega_n(A, \mathbf{P}_\bullet)$.

Syzygies help compute projective dimension.

Lemma 11.123. *The following conditions are equivalent for a left R-module A.*

(i) $pd(A) \leq n$.

(ii) $\mathrm{Ext}^k_R(A, B) = \{0\}$ *for all modules B and all $k \geq n + 1$.*

(iii) $\mathrm{Ext}^{n+1}_R(A, B) = \{0\}$ *for all modules B.*

(iv) *for every projective resolution $\mathbf{P}_\bullet$ of A, the $(n-1)$st syzygy $\Omega_{n-1}(A, \mathbf{P}_\bullet)$ is projective.*

(v) *there exists a projective resolution $\mathbf{P}_\bullet$ of A with $\Omega_{n-1}(A, \mathbf{P}_\bullet)$ projective.*

Proof. (i) ⇒ (ii).

By hypothesis, there is a projective resolution $\mathbf{P}_\bullet$ of A with $P_k = \{0\}$ for all $k \geq n + 1$. Necessarily, all the maps $d_k \colon P_k \to P_{k-1}$ are zero for $k \geq n + 1$, and so

$$\mathrm{Ext}^k_R(A, B) = \frac{\ker(d_{k+1})^*}{\mathrm{im}(d_k)^*} = \{0\}.$$

(ii) ⇒ (iii). Obvious.

(iii) ⇒ (iv).

If $\mathbf{P}_\bullet = \cdots \to P_n \to P_{n-1} \to \cdots \to P_1 \to P_0 \to A \to 0$ is a projective resolution of A, then $\mathrm{Ext}^{n+1}_R(A, B) \cong \mathrm{Ext}^1_R(\Omega_{n-1}(A, \mathbf{P}_\bullet), B)$, by Proposition 11.121. But the last group is $\{0\}$, by hypothesis, so that $\Omega_{n-1}(A)$ is projective, by Corollary 10.86.

(iv) ⇒ (v). Obvious.

(v) ⇒ (i).

If

$$\cdots \to P_n \to P_{n-1} \to \cdots \to P_1 \to P_0 \to A \to 0$$

is a projective resolution of A, then

$$0 \to \Omega_{n-1}(A) \to P_{n-1} \to \cdots \to P_1 \to P_0 \to A \to 0$$

is an exact sequence. Since $\Omega_{n-1}(A)$ is projective, the last sequence is a projective resolution of A, and so $pd(A) \leq n$. •

Example 11.124.
Let G be a finite cyclic group with $|G| > 1$. If $\mathbb{Z}$ is viewed as a trivial $\mathbb{Z}G$-module, then $pd(\mathbb{Z}) = \infty$, because Corollary 10.108 gives, for all odd n,

$$H^n(G, \mathbb{Z}) = \operatorname{Ext}^n_{\mathbb{Z}G}(\mathbb{Z}, \mathbb{Z}) \neq \{0\}. \quad ◀$$

The following definition will soon be simplified.

Definition. If R is a ring, then its ***left projective global dimension*** is defined by

$$lpD(R) = \sup\{pd(A) : A \in \operatorname{obj}({}_R\mathbf{Mod})\}.$$

Proposition 11.125. *For any ring R,*

$$lpD(R) \leq n \quad \textit{if and only if} \quad \operatorname{Ext}^{n+1}_R(A, B) = \{0\}$$

for all left R-modules A and B.

Proof. This follows at once from the equivalence of (i) and (iii) in Lemma 11.123. •

Example 11.126.
(i) A ring R is semisimple if and only if $lpD(R) = 0$. Thus, global dimension is a measure of how far a ring is from being semisimple.

(ii) A ring R is left hereditary if and only if $lpD(R) \leq 1$. In particular, a domain R is Dedekind if and only if $pD(R) \leq 1$. ◀

A similar discussion can be given using injective resolutions.

Definition. Let R be a ring and let B be a left R-module. If there is an injective resolution

$$0 \to B \to E^0 \to E^1 \to \cdots \to E^n \to 0,$$

then we write $id(B) \leq n$. If $n \geq 0$ is the smallest integer such that $id(B) \leq n$, then we say that B has ***injective dimension*** n; if there is no finite injective resolution of B, then $id(B) = \infty$.

Example 11.127.
(i) A module B is injective if and only if $id(B) = 0$. We may thus regard $id(B)$ as a measure of how far away B is from being injective.

(ii) The injective and projective dimensions of a module A can be distinct. For example, the abelian group $A = \mathbb{Z}$ has $pd(A) = 0$ and $id(A) = 1$.

(iii) If R is a Dedekind ring, then Theorem 11.101 says that every quotient module of an injective R-module is injective. Hence, if $\eta \colon B \to E$ is an imbedding of an R-module B into an injective R-module E, then

$$0 \to B \xrightarrow{\eta} E \to \operatorname{coker} \eta \to 0$$

is an injective resolution of B. It follows that $id(B) \leq 1$. ◀

Definition. Let $\mathbf{E}^\bullet = 0 \to B \xrightarrow{\eta} E^0 \xrightarrow{d^0} E^1 \xrightarrow{d^1} E^2 \to \cdots$ be an injective resolution of a module B. If $n \geq 0$, then the nth ***cosyzygy*** is

$$\mho^n(B, \mathbf{E}^\bullet) = \begin{cases} \operatorname{coker} \eta & \text{if } n = 0 \\ \operatorname{coker} d^{n-1} & \text{if } n \geq 1. \end{cases}$$

Proposition 11.128. *For every $n \geq 1$, for all left R-modules A and B, and for every injective resolution $\mathbf{E}^\bullet$ of A, there is an isomorphism*

$$\operatorname{Ext}_R^{n+1}(A, B) \cong \operatorname{Ext}_R^1(A, \mho^{n-1}(B, \mathbf{E}^\bullet)).$$

Proof. Dual to the proof of Proposition 11.121. •

Corollary 11.129. *For all left R-modules A and B, for all $n \geq 0$, and for any injective resolutions $\mathbf{E}^\bullet$ and $\mathbf{E}'^\bullet$ of B, there is an isomorphism*

$$\operatorname{Ext}_R^1(A, \mho^n(B, \mathbf{E}^\bullet)) \cong \operatorname{Ext}_R^1(A, \mho^n(B, \mathbf{E}'^\bullet)).$$

Proof. Dual to the proof of Corollary 11.122. •

Two modules $\mho$ and $\mho'$ are called ***injectively equivalent*** if there exist injective modules E and E' with $\mho \oplus E \cong \mho' \oplus E'$. Exercise 11.63 on page 983 shows that any two nth cosyzygies of a module B are injectively equivalent. We often abuse notation and speak of *the* nth cosyzygy of a module, writing $\mho^n(B)$ instead of $\mho^n(B, \mathbf{E}^\bullet)$.

Cosyzygies help compute injective dimension.

Lemma 11.130. *The following conditions are equivalent for a left R-module B.*

(i) $id(B) \leq n$.

(ii) $\operatorname{Ext}_R^k(A, B) = \{0\}$ *for all modules A and all $k \geq n + 1$.*

(iii) $\operatorname{Ext}_R^{n+1}(A, B) = \{0\}$ *for all modules A.*

(iv) *for every injective resolution $\mathbf{E}^\bullet$ of B, the $(n-1)$st cosyzygy $\mho^{n-1}(B, \mathbf{E}^\bullet)$ is injective.*

(v) *there exists an injective resolution $\mathbf{E}^\bullet$ of B with $\mho^{n-1}(B, \mathbf{E}^\bullet)$ injective.*

Proof. Dual to that of Lemma 11.123, using Exercise 10.49 on page 869 •

Definition. If R is a ring, then its ***left injective global dimension*** is defined by

$$liD(R) = \sup\{id(B) : B \in \operatorname{obj}({}_R\mathbf{Mod})\}.$$

Proposition 11.131. *For any ring R,*

$$liD(R) \leq n \text{ if and only if } \operatorname{Ext}_R^{n+1}(A, B) = \{0\}$$

for all left R-modules A and B.

Proof. This follows at once from the equivalence of (i) and (iii) in Lemma 11.130. •

Theorem 11.132. *For every ring R,*

$$lpD(R) = liD(R).$$

Proof. This follows at once from Propositions 11.125 and 11.131, for each number is equal to the smallest n for which $\mathrm{Ext}_R^{n+1}(A, B) = \{0\}$ for all left R-modules A and B. •

Definition. The ***left global dimension*** of a ring R is the common value of the left projective global dimension and the left injective global dimension:

$$lD(R) = lpD(R) = liD(R).$$

If R is commutative, then we denote its global dimension by $D(R)$

There is also a ***right global dimension*** $rD(R) = lD(R^{\mathrm{op}})$ of a ring R. If R is commutative, then $lD(R) = rD(R)$ and we write $D(R)$. Since left semisimple rings are also right semisimple, by Corollary 8.57, we have $lD(R) = 0$ if and only if $rD(R) = 0$. On the other hand, there are examples of rings in which these two dimensions differ.

We are now going to see that $lD(R)$ can be computed from cyclic left R-modules.

Lemma 11.133. *A left R-module B is injective if and only if* $\mathrm{Ext}_R^1(R/I, B) = \{0\}$ *for every left ideal I.*

Proof. If B is injective, then $\mathrm{Ext}_R^1(A, B)$ vanishes for every right R-module A. Conversely, suppose that $\mathrm{Ext}_R^1(R/I, B) = \{0\}$ for every left ideal I. Applying $\mathrm{Hom}_R(\ , B)$ to the exact sequence $0 \to I \to R \to R/I \to 0$ gives exactness of

$$\mathrm{Hom}_R(R, B) \to \mathrm{Hom}_R(I, B) \to \mathrm{Ext}_R^1(R/I, B) = 0.$$

That is, every R-map $f : I \to B$ can be extended to an R-map $R \to B$ (see Proposition 7.63). But this is precisely the Baer criterion, Theorem 7.68, and so B is injective. •

The next result says that $lD(R)$ can be computed from $pd(M)$ for finitely generated R-modules M; in fact, $lD(R)$ can even be computed from $pd(M)$ for M cyclic.

Theorem 11.134 (Auslander). *For any ring R,*

$$lD(R) = \sup\{pd(R/I) : I \text{ is a left ideal}\}.$$

Proof. (***Matlis***) If $\sup\{pd(R/I)\} = \infty$, we are done. Therefore, we may assume there is an integer $n \geq 0$ with $pd(R/I) \leq n$ for every left ideal I. By Lemma 11.130, $\mathrm{Ext}_R^{n+1}(R/I, B) = \{0\}$ for every left R-module B. But $lpD(R) = liD(R)$, by Theorem 11.132, so that it suffices to prove that $id(B) \leq n$ for every B. Let $\mathbf{E}^\bullet$ be an injective resolution of B, with $(n-1)$st cosyzygy $\mho^{n-1}(B)$. By Corollary 11.128, $\{0\} = \mathrm{Ext}_R^{n+1}(R/I, B) \cong \mathrm{Ext}_R^1(R/I, \mho^{n-1}(B))$. Now Lemma 11.133 gives $\mho^{n-1}(B)$ injective, and so Lemma 11.130 gives $id(B) \leq n$, as desired. •

This theorem explains why every ideal in a Dedekind ring being projective is such a strong condition.

Just as Ext defines the global dimension of a ring R, we can use Tor to define the *weak dimension* (or Tor-*dimension*) of a ring R.

Definition. Let R be a ring and let A be a right R-module. A ***flat resolution*** of A is an exact sequence

$$\cdots \to F_n \to \cdots \to F_1 \to F_0 \to A \to 0$$

in which each F_n is a flat right R-module.

If there is a finite flat resolution

$$0 \to F_n \to \cdots \to F_1 \to F_0 \to A \to 0,$$

then we write $fd(A) \leq n$. If $n \geq 0$ is the smallest integer such that $fd(A) \leq n$, then we say that A has ***flat dimension*** n; if there is no finite flat resolution of A, then $fd(A) = \infty$.

Example 11.135.

(i) A module A is flat if and only if $fd(A) = 0$. We may thus regard $fd(A)$ as a measure of how far away A is from being flat.

(ii) Since projective modules are flat, every projective resolution of A is a flat resolution. It follows that if R is any ring, then $fd(A) \leq pd(A)$ for every R-module A.

(iii) If R is a Dedekind ring and A is an R-module, then $fd(A) \leq pd(A) \leq 1$, by (ii). Corollary 11.109 says that every torsion-free R-module is flat (the converse is true as well). Hence, $fd(A) = 1$ if and only if A is not torsion-free. ◂

Definition. Let $\mathbf{F}_\bullet = \cdots \to F_2 \xrightarrow{d_2} F_1 \xrightarrow{d_1} F_0 \xrightarrow{\varepsilon} A \to 0$ be a flat resolution of a module A. If $n \geq 0$, then the nth ***yoke*** is

$$Y_n(A, \mathbf{F}_\bullet) = \begin{cases} \ker \varepsilon & \text{if } n = 0 \\ \ker d_n & \text{if } n \geq 1. \end{cases}$$

The term *yoke* is not standard; it is a translation of the Greek $\sigma\upsilon\zeta\upsilon\gamma\acute{\iota}\alpha$ (syzygy).

Proposition 11.136. *For every $n \geq 1$, for all right R-modules A and left R-modules B, and for every flat resolution $\mathbf{F}_\bullet$ of A, there is an isomorphism*

$$\operatorname{Tor}_{n+1}^R(A, B) \cong \operatorname{Tor}_1^R(A, Y_{n-1}(B, \mathbf{F}_\bullet)).$$

Proof. Dual to the proof of Proposition 11.121. •

Corollary 11.137. *For every right R-module A and left R-module B, for all $n \geq 0$, and for any flat resolutions $\mathbf{F}_\bullet$ and $\mathbf{F}'_\bullet$ of B, there is an isomorphism*

$$\operatorname{Tor}_1^R(A, Y_n(B, \mathbf{F}_\bullet)) \cong \operatorname{Tor}_1^R(A, Y_n(B, \mathbf{F}'_\bullet)).$$

Proof. Dual to the proof of Corollary 11.122. •

Lemma 11.138. *The following conditions are equivalent for a right R-module A.*

(i) $fd(A) \le n$.

(ii) $\operatorname{Tor}_k^R(A, B) = \{0\}$ *for all* $k \ge n+1$ *and all left R-modules B.*

(iii) $\operatorname{Tor}_{n+1}^R(A, B) = \{0\}$ *for all left R-modules B.*

(iv) *For every flat resolution* $\mathbf{F}^\bullet$ *of A, the* $(n-1)$*st yoke* $Y_{n-1}(A, \mathbf{F}^\bullet)$ *is flat.*

(v) *There exists a flat resolution* $\mathbf{F}^\bullet$ *of A with flat* $(n-1)$*st yoke* $Y_{n-1}(A, \mathbf{F}^\bullet)$.

Proof. As the proof of Lemma 11.123. •

Definition. The ***right weak dimension*** of a ring R is defined by

$$rwD(R) = \sup\{fd(A) : A \in \operatorname{obj}(\mathbf{Mod}_R)\}.$$

Proposition 11.139. *For any ring R, $rwD(R) \le n$ if and only if* $\operatorname{Tor}_{n+1}^R(A, B) = \{0\}$ *for every left R-module B.*

Proof. This follows at once from Lemma 11.138. •

We define the flat dimension of left R-modules in the obvious way.

Definition. The ***left weak dimension*** of a ring R is defined by

$$lwD(R) = \sup\{fd(B) : B \in \operatorname{obj}({}_R\mathbf{Mod})\}.$$

Theorem 11.140. *For any ring R,*

$$rwD(R) = lwD(R).$$

Proof. If either dimension is finite, then the left or right weak dimension is the smallest $n \ge 0$ with $\operatorname{Tor}_{n+1}^R(A, B) = \{0\}$ for all right R-modules A and all left R-modules B. •

Definition. The ***weak dimension*** of a ring R, denoted by $wD(R)$, is the common value of $rwD(R)$ and $lwD(R)$.

As we have remarked earlier, there are (noncommutative) rings whose left global dimension and right global dimension can be distinct. In contrast, weak dimension has no left/right distinction, because tensor and Tor involve both left and right modules simultaneously.

Example 11.141.
A ring R has $wD(R) = 0$ if and only if every module is flat. These rings turn out to be ***von Neumann regular***: for each $a \in R$, there exists $a' \in R$ with $aa'a = a$. Examples of such rings are Boolean rings (rings R in which $r^2 = r$ for all $r \in R$), and $\operatorname{End}_k(V)$, where V is a (possibly infinite-dimensional) vector space over a field k. See Rotman, *An Introduction to Homological Algebra*, pages 119–120. ◂

The next proposition explains why weak dimension is so called.

Proposition 11.142. *For any ring R,*

$$wD(R) \le \min\{lD(R), rD(R)\}.$$

Proof. It suffices to prove that $fd(A) \le pd(A)$ for any right R-module A. If $pd(A) = \infty$, there is nothing to prove; if $pd(A) \le n$, there is a projective resolution

$$0 \to P_n \to \cdots \to P_0 \to A \to 0.$$

Since every projective module is flat, this is a flat resolution showing that $fd(A) \le n$. Hence, $wD(R) \le rD(R)$. A similar argument shows that $wD(R) \le lD(R)$. •

Corollary 11.143. *Suppose that* $\text{Ext}^n_R(A, B) = \{0\}$ *for all left R-modules A and B. Then* $\text{Tor}^R_n(C, D) = \{0\}$ *for all right R-modules C and all left R-modules D.*

Proof. If $\text{Ext}^n_R(A, B) = \{0\}$ for all A, B, then $lD(R) \le n - 1$; if $\text{Tor}^R_n(C, D) \ne \{0\}$ for some C, D, then $n \le wD(DR)$. This contradicts Proposition 11.142:

$$lD(R) \le n - 1 < n \le wD(R). \quad \bullet$$

Lemma 11.144. *A left R-module B is flat if and only if* $\text{Tor}^R_1(R/I, B) = \{0\}$ *for every right ideal I.*

Proof. Exactness of $0 \to I \xrightarrow{i} R \to R/I \to 0$ gives exactness of

$$0 = \text{Tor}^R_1(R, B) \to \text{Tor}^R_1(R/I, B) \to I \otimes_R B \xrightarrow{i \otimes 1} R \otimes_R B.$$

Therefore, $i \otimes 1$ is an injection if and only if $\text{Tor}^R_1(R/I, B) = \{0\}$. On the other hand, B is flat if and only if $i \otimes 1$ is an injection for every right ideal I, by Corollary 8.108. •

As global dimension, weak dimension can be computed from cyclic modules.

Corollary 11.145. *For any ring R,*

$$\begin{aligned} wd(R) &= \sup\{fd(R/I) : I \text{ is a right ideal of } R\} \\ &= \sup\{fd(R/J) : J \text{ is a left ideal of } R\}. \end{aligned}$$

Proof. This proof is similar to the proof of Theorem 11.134, using Lemma 11.144 instead of Lemma 11.133. •

Theorem 11.146. *Let R be a left noetherian ring.*

(i) *If A is a finitely generated left R-module, then*

$$pd(A) = fd(A).$$

(ii)

$$wD(R) = lD(R).$$

In particular, if R is a commutative noetherian ring, then

$$wD(R) = D(R).$$

Proof. (i) It is always true that $fd(A) \leq pd(A)$, for every projective resolution is a flat resolution. For the reverse inequality, it is enough to prove that if $fd(A) \leq n$, then $pd(A) \leq n$. By Lemma 11.37, there is a projective resolution of A,

$$\cdots \to P_n \to \cdots \to P_0 \to A \to 0,$$

in which each P_i is finitely generated. Now this is also a flat resolution, so that, by Lemma 11.138, $fd(A) \leq n$ implies $Y_n = \ker(P_{n-1} \to P_{n-2})$ is flat. But every finitely generated flat left R-module is projective, by Corollary 8.111 (because R is left noetherian), and so

$$0 \to Y_{n-1} \to P_{n-1} \to \cdots \to P_0 \to A \to 0$$

is a projective resolution. Therefore, $pd(A) \leq n$.

(ii) By Theorem 11.134, we have $lD(R)$ is the supremum of projective dimensions of cyclic left R-modules, and by Corollary 11.145, we have $wd(R)$ is the supremum of flat dimensions of cyclic left R-modules. But part (i) gives $fd(A) = pd(A)$ for every finitely generated right R-module A, and this suffices to prove the result. •

We are now going to compute the global dimension $D(k[x_1, \ldots, x_n])$ of a polynomial ring over a field (the result is called *Hilbert's theorem on syzygies*).

Lemma 11.147. *If $0 \to A' \to A \to A'' \to 0$ is a short exact sequence, then*

$$pd(A'') \leq 1 + \max\{pd(A), pd(A')\}.$$

Proof. We may assume the right side is finite, or there is nothing to prove; let $pd(A) \leq n$ and $pd(A') \leq n$. Applying $\text{Hom}(\ , B)$, where B is any module, to the short exact sequence gives the long exact sequence

$$\cdots \to \text{Ext}^{n+1}(A', B) \to \text{Ext}^{n+2}(A'', B) \to \text{Ext}^{n+2}(A, B) \to \cdots .$$

The two outside terms are $\{0\}$, by Lemma 11.123(ii), so that exactness forces $\text{Ext}^{n+2}(A'', B) = \{0\}$ for all B. The same lemma gives $pd(A'') \leq n + 1$. •

We wish to compare global dimension of R and $R[x]$, and so we consider the $R[x]$-module $R[x] \otimes_R M \cong M[x]$ arising from an R-module M. In Chapter 9, on page 684, we called this module $M[x]$.

Definition. If M is an R-module over a commutative ring R, define

$$M[x] = \sum_{i \geq 0} M_i,$$

where $M_i \cong M$ for all i. The R-module $M[x]$ is an $R[x]$-module if we define

$$x\Big(\sum_i x^i m_i\Big) = \sum_i x^{i+1} m_i.$$

In Lemma 9.55, we proved that if V is a free R-module over a commutative ring R, then $V[x]$ is a free $R[x]$-module. The next result generalizes this from $pd(V) = 0$ to higher dimensions.

Lemma 11.148. *For every R-module M, where R is a commutative ring,*

$$pd_R(M) = pd_{R[x]}(M[x]).$$

Proof. It suffices to prove that if one of the dimensions is finite and at most n, then so is the other.

If $pd(M) \leq n$, then there is an R-projective resolution

$$0 \to P_n \to \cdots \to P_0 \to M \to 0.$$

Since $R[x]$ is a free R-module, it is a flat R-module, and so there is an exact sequence of $R[x]$-modules

$$0 \to R[x] \otimes_R P_n \to \cdots \to R[x] \otimes_R P_0 \to R[x] \otimes_R M \to 0.$$

But $R[x] \otimes_R M \cong M[x]$ and $R[x] \otimes_R P_n$ is a projective $R[x]$-module (for a projective is a direct summand of a free module). Therefore, $pd_{R[x]}(M[x]) \leq n$.

If $pd(M[x]) \leq n$, then there is an $R[x]$-projective resolution

$$0 \to Q_n \to \cdots \to Q_0 \to M[x] \to 0.$$

As an R-module, $M[x] \cong \sum_{n \geq 1} M_n$, where $M_n \cong M$ for all n. By Exercise 11.69 on page 984, $pd_R(M[x]) = pd_R(M)$. Each projective $R[x]$-module Q_i is an $R[x]$-summand of a free $R[x]$-module F_i; a fortiori, Q_i is an R-direct summand of F_i. But $R[x]$ is a free R-module, so that F_i is also a free R-module. Therefore, as an R-module, Q_i is projective, and so $pd_R(M) \leq pd_{R[x]}(M[x])$. •

Corollary 11.149. *If R is a commutative ring and $D(R) = \infty$, then $D(R[x]) = \infty$.*

Proof. If $D(R) = \infty$, then for every integer n, there exists an R-module M_n with $n < pd(M_n)$. By the lemma, $n < pd_{R[x]}(M_n[x])$, and so $D(R[x]) = \infty$. •

Proposition 11.150. *For every commutative ring R,*

$$D(R[x]) \leq D(R) + 1.$$

Proof. Recall the characteristic sequence, Theorem 9.56: If M is an R-module and $T : M \to M$ is an R-map, then there is an exact sequence of $R[x]$-modules

$$0 \to M[x] \to M[x] \to M^T \to 0,$$

where M^T is the $R[x]$-module M with scalar multiplication given by $ax^i m = aT^i(m)$. If M is already an $R[x]$-module and $T : M \to M$ is the R-map $m \mapsto xm$, then $M^T = M$. By Lemma 11.147,

$$\begin{aligned} pd_{R[x]}(M) &\leq 1 + pd_{R[x]}(M[x]) \\ &= 1 + pd_R(M) \\ &\leq 1 + D(R). \quad \bullet \end{aligned}$$

We proceed to prove the reverse inequality.

Definition. If M is an R-module, where R is a commutative ring, then an element $c \in R$ is ***regular*** on M (or is M-***regular***) if the R-map $M \to M$, given by $m \mapsto cm$, is injective. Otherwise, c is a ***zero divisor*** on M; that is, there is some nonzero $m \in M$ with $cm = 0$.

Before stating the next theorem, let us explain the notation. Suppose that R is a commutative ring, $c \in R$, and $R^* = R/Rc$. If M is an R-module, then M/cM is an (R/Rc)-module; that is, M/cM is a R^*-module. On the other hand, every R^*-module A^* can also be viewed as an R-module. If $\sigma : (R/Rc) \times A^* \to A^*$ is the given scalar multiplication and if $\nu : R \to R/Rc$ is the natural map, then $\sigma(\nu \times 1_{A^*}) : R \times A^* \to A^*$ is a scalar multiplication. In more down-to-earth language, if $r^* = r + Rc$, then

$$r^* a = ra.$$

We denote A^* viewed in this way as an R-module by $A^\flat$. Exercise 11.72 on page 984 asks you to prove that $M^* \cong (R/cR) \otimes_R M$ and $A^\flat \cong \mathrm{Hom}_R(R/cR, A^*)$, so that these constructions involve an adjoint pair of functors.

Proposition 11.151 (Rees Lemma). *Let R be a commutative ring, let $c \in R$ be neither a unit nor a zero divisor, and let $R^* = R/Rc$. If c is regular on an R-module M, then there are natural isomorphisms, for every R^*-module A^* and all $n \geq 0$,*

$$\mathrm{Ext}^n_{R^*}(A^*, M/cM) \cong \mathrm{Ext}^{n+1}_R(A^\flat, M),$$

where $A^\flat$ is the R^*-module A^* viewed as an R-module.

Proof. Recall Theorem 10.45, the axioms characterizating Ext functors. Given a sequence of contravariant functors $G^n\colon {}_{R^*}\mathbf{Mod} \to \mathbf{Ab}$, for $n \geq 0$, such that:

(i) for every short exact sequence $0 \to A^* \to B^* \to C^* \to 0$ of R^*-modules, there is a long exact sequence with natural connecting homomorphisms

$$\cdots \to G^n(C^*) \to G^n(B^*) \to G^n(A^*) \to G^{n+1}(C^*) \to \cdots;$$

(ii) G^0 and $\mathrm{Hom}_{R^*}(\ , L^*)$ are naturally equivalent, for some R^*-module L^*;

(iii) $G^n(P^*) = 0$ for all projective R^*-modules P^* and all $n \geq 1$;

then G^n is naturally equivalent to $\mathrm{Ext}^n_{R^*}(\ , L^*)$ for all $n \geq 0$.

Define contravariant functors $G^n\colon {}_{R^*}\mathbf{Mod} \to \mathbf{Ab}$ by $G^n = \mathrm{Ext}^{n+1}_R(\ ^\flat, M)$. Thus, for all R^*-modules A^*,

$$G^n(A^*) = \mathrm{Ext}^{n+1}_R(A^\flat, M).$$

Since Axiom (i) holds for the functors Ext^n, it also holds for the functors G^n. Let us prove Axiom (ii). The map $\mu_c\colon M \to M$, defined by $m \mapsto cm$, is an injection, because c is M-regular, and so the sequence $0 \to M \xrightarrow{\mu_c} M \to M/cM \to 0$ is exact. Consider the portion of the long exact sequence, where A^* is an R^*-module:

$$\mathrm{Hom}_R(A^\flat, M) \to \mathrm{Hom}_R(A^\flat, M/cM) \xrightarrow{\partial} \mathrm{Ext}^1_R(A^\flat, M) \xrightarrow{(\mu_c)_*} \mathrm{Ext}^1_R(A^\flat, M).$$

We claim that ∂ is an isomorphism. If $a \in A^\flat$, then $ca = 0$, because A^* is an R^*-module (remember that $R^* = R/cR$). Hence, if $f \in \mathrm{Hom}_R(A^\flat, M)$, then $cf(a) = f(ca) = f(0) = 0$. Since $\mu_c\colon M \to M$ is an injection, $f(a) = 0$ for all $a \in A^\flat$. Thus, $f = 0$, $\ker\partial = \mathrm{Hom}_R(A^\flat, M) = \{0\}$, and ∂ is an injection. The map $(\mu_c)_*\colon \mathrm{Ext}^1_R(A^\flat, M) \to \mathrm{Ext}^1_R(A^\flat, M)$ is multiplication by c, by Example 10.60. On the other hand, Example 10.70 shows that if $\mu'_c\colon A^\flat \to A^\flat$ is multiplication by c, then the induced map $(\mu'_c)^*$ on Ext is also multiplication by c. But $\mu'_c = 0$, because A^* is a (R/cR)-module, and so $(\mu'_c)^* = 0$. Hence, $(\mu_c)_* = (\mu'_c)^* = 0$. Therefore, $\mathrm{im}\,\partial = \ker(\mu_c)_* = \mathrm{Ext}^1_R(A^\flat, M)$, and so ∂ is a surjection. It follows that

$$\partial\colon \mathrm{Hom}_R(A^\flat, M/cM) \to \mathrm{Ext}^1_R(A^\flat, M)$$

is an isomorphism, natural because it is the connecting homomorphism. By Exercise 11.70 on page 984, there is a natural isomorphism

$$\mathrm{Hom}_{R^*}(A^*, M/cM) \to \mathrm{Hom}_R(A^\flat, M/cM).$$

The composite

$$\mathrm{Hom}_{R^*}(A^*, M/cM) \to \mathrm{Hom}_R(A^\flat, M/cM) \to \mathrm{Ext}^1_R(A^\flat, M) = G^0(A^*)$$

is a natural isomorphism; hence, its inverse defines a natural equivalence

$$G^0 \to \mathrm{Hom}_{R^*}(\ , M/cM).$$

Setting $L^* = M/cM$ completes the verification of Axiom (ii).

It remains to verify Axiom (iii): $G^n(P^*) = \{0\}$ whenever P^* is a projective R^*-module and $n \geq 1$. In fact, since G^n is an additive functor and since every projective is a summand of a free module, we may assume that P^* is a free R^*-module with basis, say, E. If $Q = \sum_{e \in E} Re$ is the free R-module with basis E, then there is an exact sequence of R-modules

$$0 \to Q \xrightarrow{\mu_c} Q \to P^* \to 0. \tag{6}$$

The first arrow is an injection because c is not a zero divisor in R; the last arrow is a surjection because

$$\begin{aligned} Q/cQ &= \big(\sum_{e \in E} Re\big)/\big(\sum_{e \in E} Rce\big) \\ &\cong \sum_{e \in E} (R/Rc)e = \sum R^* e = P^*. \end{aligned}$$

The long exact sequence arising from (6) is

$$\cdots \to \operatorname{Ext}_R^n(Q, M) \to \operatorname{Ext}_R^{n+1}(P^\flat, M) \to \operatorname{Ext}_R^{n+1}(Q, M) \to \cdots .$$

Since Q is R-free and $n \geq 1$, the outside terms are $\{0\}$, and exactness gives $G^n(P^*) = \operatorname{Ext}_R^{n+1}(P^\flat, M) = \{0\}$. Therefore,

$$\operatorname{Ext}_R^{n+1}(A^\flat, M) = G^n(A^*) \cong \operatorname{Ext}_{R^*}^n(A^*, M/cM). \quad \bullet$$

Theorem 11.152. *For every commutative ring k,*

$$D(k[x]) = D(k) + 1.$$

Proof. We have proved $D(k[x]) \leq D(k) + 1$ in Proposition 11.150, and so it suffices to prove the reverse inequality.

In the notation of the Rees lemma, Proposition 11.151, let us write $R = k[x]$, $c = x$, and $R^* = k$. Let A be a k-module with $pd(A) = n$. By Exercise 11.65 on page 984, there is a free k-module F with $\operatorname{Ext}_k^n(A, F) \neq \{0\}$; of course, multiplication by x is an injection $F \to F$. As in the proof of the Rees lemma, there is a free $k[x]$-module $Q = k[x] \otimes_k F$ with $Q/xQ \cong F$. The Rees lemma gives

$$\operatorname{Ext}_{k[x]}^{n+1}(A, Q) \cong \operatorname{Ext}_k^n(A, Q/xQ) \cong \operatorname{Ext}_k^n(A, F) \neq \{0\}$$

(A is viewed as a $k[x]$-module via $k[x] \to k$). Therefore, $pd_{k[x]}(A) \geq n + 1$, and so $D(k[x]) \geq n + 1 = D(k) + 1$. $\bullet$

Corollary 11.153 (Hilbert's Theorem on Syzygies). *If k is a field, then*

$$D(k[x_1, \ldots, x_n]) = n.$$

Proof. Since $D(k) = 0$ and $D(k[x]) = 1$ for every field k, the result follows from Theorem 11.152 by induction on $n \geq 0$. •

Hilbert's theorem on syzygies implies that if $R = k[x_1, \ldots, x_n]$, where k is a field, then every finitely generated R-module M has has a resolution

$$0 \to P_n \to P_{n-1} \to \cdots \to P_0 \to M \to 0,$$

where P_i is free for all $i < n$ and P_n is projective. We say that a (necessarily finitely generated) R-module M, over an arbitrary commutative ring R, has ***FFR*** (*finite free resolution*) if it has a resolution in which every P_i, including the last P_n, is a finitely generated free module. Hilbert's theorem on syzygies can be improved to the theorem that if k is a field, then every finitely generated $k[x_1, \ldots, x_n]$-module has FFR (see Kaplansky, *Commutative Rings*, page 134). (Of course, this result also follows from the more difficult Quillen–Suslin theorem, which says that every projective $k[x_1, \ldots, x_n]$-module, where k is a field, is free.)

EXERCISES

11.62 (i) If $A \to B \xrightarrow{f} C \to D$ is an exact sequence, and if X is any module, prove that there is an exact sequence

$$A \to B \oplus X \xrightarrow{f \oplus 1_X} C \oplus X \to D.$$

(ii) Let

$$\mathbf{P}_\bullet = \cdots \to P_2 \xrightarrow{d_2} P_1 \xrightarrow{d_1} P_0 \xrightarrow{\varepsilon} A \to 0$$

and

$$\mathbf{P}'_\bullet = \cdots \to P'_2 \xrightarrow{d'_2} P'_1 \xrightarrow{d'_1} P'_0 \xrightarrow{\varepsilon'} A \to 0$$

be projective resolutions of a left R-module A. For all $n \geq 0$, prove that there are projective modules Q_n and Q'_n with

$$\Omega_n(A, \mathbf{P}_\bullet) \oplus Q_n \cong \Omega_n(A, \mathbf{P}'_\bullet) \oplus Q'_n.$$

Hint. Proceed by induction on $n \geq 0$, using Schanuel's lemma, Proposition 7.60.

11.63 Let

$$0 \to B \to E^0 \to E^1 \to E^2 \to \cdots$$

and

$$0 \to B \to E'^0 \to E'^1 \to E'^2 \to \cdots$$

be injective resolutions of a left R-module B. For all $n \geq 0$, prove that there are injective modules I_n and I'_n with

$$\mho^n(B, \mathbf{E}^\bullet) \oplus I_n \cong \mho^n(B, \mathbf{E}'^\bullet) \oplus I'_n.$$

Hint. The proof is dual to that of Exercise 11.62.

11.64 Show that there are flat resolutions

$$0 \to \mathbb{Z} \to \mathbb{Q} \to \mathbb{Q}/\mathbb{Z} \to 0$$

and

$$0 \to K \to F \to \mathbb{Q}/\mathbb{Z} \to 0,$$

where F is free abelian, but that $\mathbb{Z} \oplus F \not\cong \mathbb{Q} \oplus K$.

11.65 If A is an R-module with $pd(A) = n$, prove that there exists a free R-module F with $\mathrm{Ext}^n_R(A, F) \neq \{0\}$.
Hint. Every module is a quotient of a free module.

11.66 If G is a finite cyclic group of order not 1, prove that

$$lD(\mathbb{Z}G) = \infty = rD(\mathbb{Z}G).$$

Hint. Use Theorem 10.107.

11.67 (***Auslander***) If R is both left noetherian and right noetherian, prove that

$$lD(R) = rD(R).$$

Hint. Use weak dimension.

11.68 Prove that a noetherian von Neumann regular ring is semisimple.
Hint. See Example 11.141.

11.69 If $\{M_\alpha : \alpha \in A\}$ is a family of left R-modules, prove that

$$pd\Big(\sum_{\alpha \in A} M_\alpha\Big) = \sup_{\alpha \in A}\big\{pd(M_\alpha)\big\}.$$

11.70 If $\varphi\colon R \to R^*$ is a ring homomorphism and A^* and B^* are R^*-modules, prove that there is a natural isomorphism

$$\mathrm{Hom}_{R^*}(A^*, B^*) \to \mathrm{Hom}_R(A^\flat, B^\flat),$$

where $A^\flat$ is A^* viewed as an R-module.

11.71 (i) If $0 \to A' \to A \to A'' \to 0$ is an exact sequence, prove that

$$pd(A) \le \max\{pd(A'), pd(A'')\}.$$

(ii) If the sequence in part (i) is not split and if $pd(A') = pd(A'') + 1$, prove that

$$pd(A) = \max\{pd(A'), pd(A'')\}.$$

11.72 Let k be a commutative ring, let $c \in k$, and let $k^* = k/ck$.

(i) If M is a k-module, define $M^* = M/cM$. Prove that $M^* \cong (k/ck) \otimes_k M$.

(ii) If A^* is a k^*-module, define $A^\flat$ to be A^* viewed as a K-module, as on page 980. Prove that $A^\flat \cong \mathrm{Hom}_k(k/ck, A^*)$. Conclude that $M \mapsto M^*$ and $A^* \mapsto A^\flat$ form an adjoint pair of functors.

11.73 Let $\varphi\colon k \to k^*$ be a ring homomorphism.

(i) Prove that k^* is a (k^*, k)-bimodule.

(ii) Prove that every k^*-module A^* can be viewed as a k-module, denoted by $A^\flat$, and that $A^* \mapsto A^\flat$ gives an exact functor $U\colon {}_{k^*}\mathbf{Mod} \to {}_k\mathbf{Mod}$.

(iii) Prove that if $F = \text{Hom}_k(k^*, \ ) : {}_k\mathbf{Mod} \to {}_{k^*}\mathbf{Mod}$, then (U, F) and (F, U) are adjoint pairs of functors. (These functors are called ***change of rings*** functors.) Conclude that U and F preserve all direct limits and all inverse limits.

11.74 Let R be a commutative ring with FFR. Prove that every finitely generated projective R-module P has a free complement; that is, there is a finitely generated free R-module F such that $P \oplus F$ is a free R-module.

11.4 REGULAR LOCAL RINGS

We are now going to focus on (commutative) noetherian local rings, the main results being that such a ring has finite global dimension if and only if it is a *regular local* ring (regular local rings arise quite naturally in algebraic geometry), and that they are UFDs. Let us begin with a localization result.

Proposition 11.154. *Let R be a commutative noetherian ring.*

(i) *If A is a finitely generated R-module, then*
$$pd(A) = \sup_{\mathfrak{m}} pd(A_{\mathfrak{m}}),$$
where $\mathfrak{m}$ ranges over all the maximal ideals of R.

(ii)
$$D(R) = \sup_{\mathfrak{m}} D(R_{\mathfrak{m}}),$$
where $\mathfrak{m}$ ranges over all the maximal ideals of R.

Proof. (i) We first prove that $pd(A) \geq pd(A_{\mathfrak{m}})$ for every maximal ideal $\mathfrak{m}$. If $pd(A) = \infty$, there is nothing to prove, and so we may assume that $pd(A) = n < \infty$. Thus, there is a projective resolution
$$0 \to P_n \to P_{n-1} \to \cdots \to P_0 \to A \to 0.$$
Since $R_{\mathfrak{m}}$ is a flat R-module, by Theorem 11.28,
$$0 \to R_{\mathfrak{m}} \otimes P_n \to R_{\mathfrak{m}} \otimes P_{n-1} \to \cdots \to R_{\mathfrak{m}} \otimes P_0 \to A_{\mathfrak{m}} \to 0$$
is a projective resolution of $A_{\mathfrak{m}}$, and so $pd(A_{\mathfrak{m}}) \leq n$. (This implication does not need the hypotheses that R be noetherian or that A be finitely generated.)

For the reverse inequality, it suffices to assume that $\sup_{\mathfrak{m}} pd(A_{\mathfrak{m}}) = n < \infty$. Since R is noetherian, Theorem 11.146(i) says that $pd(A) = fd(A)$. Now $pd(A_{\mathfrak{m}}) \leq n$ if and only if $\text{Tor}^{R_{\mathfrak{m}}}_{n+1}(A_{\mathfrak{m}}, B_{\mathfrak{m}}) = \{0\}$ for all $R_{\mathfrak{m}}$-modules $B_{\mathfrak{m}}$, by Lemma 11.138. However, Proposition 11.35 gives an isomorphism $\text{Tor}^{R_{\mathfrak{m}}}_{n+1}(A_{\mathfrak{m}}, B_{\mathfrak{m}}) \cong \big(\text{Tor}^R_{n+1}(A, B)\big)_{\mathfrak{m}}$. Therefore, $\text{Tor}^R_{n+1}(A, B) = \{0\}$, by Proposition 11.31(i). We conclude that $n \geq pd(A)$.

(ii) This follows at once from part (i), for $D(R) = \sup_A\{pd(A)\}$, where A ranges over all finitely generated (even cyclic) R-modules, by Theorem 11.134. •

We now set up notation that will be used in the rest of this section.

Notation. We denote a commutative noetherian local ring by R, by $(R, \mathfrak{m})$, or by $(R, \mathfrak{m}, k)$, where $\mathfrak{m}$ is its unique maximal ideal and k is its ***residue field*** $k = R/\mathfrak{m}$.

Theorem 11.134 allows us to compute global dimension as the supremum of projective dimensions of cyclic modules. When R is a local ring, there is a dramatic improvement; global dimension is determined by the projective dimension of one cyclic module: the residue field k.

Lemma 11.155. *Let $(R, \mathfrak{m})$ be a local ring with residue field k. If A is a finitely generated R-module, then*

$$pd(A) \le n \quad \textit{if and only if} \quad \operatorname{Tor}^R_{n+1}(A, k) = \{0\}$$

Proof. Assume $pd(A) \le n$. By Example 11.135(ii), we have $fd(A) \le pd(A)$, so that $\operatorname{Tor}^R_{n+1}(A, B) = \{0\}$ for every R-module B. In particular, $\operatorname{Tor}^R_{n+1}(A, k) = \{0\}$.

We prove the converse by induction on $n \ge 0$. For the base step $n = 0$, we must prove that $\operatorname{Tor}^R_1(A, k) = \{0\}$ implies $pd(A) = 0$; that is, A is projective (hence free, since R is local). Let $\{a_1, \ldots, a_r\}$ be a minimal set of generators of A (that is, no proper subset generates A), let F be the free R-module with basis $\{e_1, \ldots, e_r\}$, and let $\varphi \colon F \to A$ be the R-map with $\varphi(e_i) = a_i$. There is an exact sequence

$$0 \to N \xrightarrow{i} F \xrightarrow{\varphi} A \to 0$$

where $N = \ker\varphi$ and i is the inclusion; as in the proof of Proposition 11.23,

$$N \subseteq \mathfrak{m}F.$$

Since $\operatorname{Tor}^R_1(A, k) = \{0\}$, the sequence

$$0 \to N \otimes_R k \xrightarrow{i\otimes 1} F \otimes_R k \xrightarrow{\varphi\otimes 1} A \otimes_R k \to 0$$

is exact. Tensor $0 \to \mathfrak{m} \to R \to k \to 0$ by N; right exactness gives a natural isomorphism

$$\tau_N \colon N \otimes_R k \to N/\mathfrak{m}N;$$

if $n \in N$ and $b \in k$, then $\tau_N : n \otimes b \mapsto n + \mathfrak{m}N$. There is a commutative diagram

$$\begin{array}{ccccc} 0 & \longrightarrow & N \otimes_R k & \xrightarrow{i\otimes 1} & F \otimes_R k \\ & & \Big\downarrow{\scriptstyle \tau_N} & & \Big\downarrow{\scriptstyle \tau_F} \\ & & N/\mathfrak{m}N & \xrightarrow[\bar{i}]{} & F/\mathfrak{m}F, \end{array}$$

where $\bar{i} : n + \mathfrak{m}N \mapsto n + \mathfrak{m}F$. Since $i \otimes 1$ is an injection, so is $\bar{i}$. But $N \subseteq \mathfrak{m}F$ says that the map $\bar{i}$ is the zero map. Therefore, $N/\mathfrak{m}N = \{0\}$, so that $N = \mathfrak{m}N$. By Nakayama's lemma, Corollary 8.32, $N = \{0\}$, and so $\varphi \colon F \to A$ is an isomorphism; that is, A is free.

For the inductive step, we must prove that if $\text{Tor}_{n+2}^R(A, k) = \{0\}$, then $pd(A) \le n + 1$. Take a projective resolution $\mathbf{P}_\bullet$ of A, and let $\Omega_n(A, \mathbf{P}_\bullet)$ be its nth syzygy. Since $\mathbf{P}_\bullet$ must also be a flat resolution of A, we have $Y_n(A, \mathbf{P}_\bullet) = \Omega_n(A, \mathbf{P}_\bullet)$. By Proposition 11.136, $\text{Tor}_{n+2}^R(A, k) \cong \text{Tor}_1^R(Y_n(A, \mathbf{P}_\bullet), k)$. The base step shows that $Y_n(A, \mathbf{P}_\bullet) = \Omega_n(A, \mathbf{P}_\bullet)$ is free, and this gives $pd(A) \le n + 1$, by Lemma 11.123. •

Corollary 11.156. *Let $(R, \mathfrak{m})$ be a local ring with residue field k. If A is a finitely generated R-module, then*

$$pd(A) = \sup\{i : \text{Tor}_i^R(A, k) \ne \{0\}\}.$$

Proof. Let $n = \sup\{i : \text{Tor}_i^R(A, k) \ne \{0\}\}$. Then $pd(A) \le n - 1$, but $pd(A) \not\le n$; that is, $pd(A) = n$. •

Theorem 11.157. *Let R be a local ring with residue field k.*

(i)
$$D(R) \le n \quad \textit{if and only if} \quad \text{Tor}_{n+1}^R(k, k) = \{0\}.$$

(ii)
$$D(R) = pd(k).$$

Proof. (i) If $D(R) \le n$, then Lemma 11.155 applies at once to give $\text{Tor}_{n+1}^R(k, k) = \{0\}$. Conversely, if $\text{Tor}_{n+1}^R(k, k) = \{0\}$, the same lemma gives $pd(k) \le n$. By Lemma 11.138, we have $\text{Tor}_{n+1}^R(A, k) = \{0\}$ for every R-module A. In particular, if A is finitely generated, then Lemma 11.155 gives $pd(A) \le n$. Finally, Theorem 11.134 shows that $D(R) = \sup_A\{pd(A)\}$, where A ranges over all finitely generated (even cyclic) R-modules. Therefore, $D(R) \le n$.

(ii) Immediate from part (i). •

Definition. A ***prime chain*** of ***length*** n in a commutative ring R is a strictly decreasing chain of prime ideals

$$\mathfrak{p}_0 \supsetneq \mathfrak{p}_1 \supsetneq \cdots \supsetneq \mathfrak{p}_n.$$

The ***height*** $\text{ht}(\mathfrak{p})$ of a prime ideal $\mathfrak{p}$ is the length of the longest prime chain with $\mathfrak{p} = \mathfrak{p}_0$. Thus, $\text{ht}(\mathfrak{p}) \le \infty$.

Example 11.158.
(i) If $\mathfrak{p}$ is a prime ideal, then $\text{ht}(\mathfrak{p}) = 0$ if and only if $\mathfrak{p}$ is a minimal prime ideal. If R is a domain, then $\text{ht}(\mathfrak{p}) = 0$ if and only if $\mathfrak{p} = \{0\}$.

(ii) If R is a Dedekind ring and $\mathfrak{p}$ is a nonzero prime ideal in R, then $\text{ht}(\mathfrak{p}) = 1$.

(iii) Let k be a field and let $R = k[X]$ be the polynomial ring in infinitely many variables $X = \{x_1, x_2, \ldots\}$. If $\mathfrak{p}_i = (x_i, x_{i+1}, \ldots)$, then $\mathfrak{p}_i$ is a prime ideal ($R/\mathfrak{p}_i \cong k[x_1, \ldots, x_{i-1}]$ is a domain) and, for every $n \geq 1$,

$$\mathfrak{p}_1 \supsetneq \mathfrak{p}_2 \supsetneq \cdots \supsetneq \mathfrak{p}_{n+1}$$

is a prime chain of length n. It follows that $\mathrm{ht}(\mathfrak{p}_1) = \infty$. ◀

Definition. If R is a commutative ring, then its ***Krull dimension*** is

$$\dim(R) = \sup\{\mathrm{ht}(\mathfrak{p}) : \mathfrak{p} \in \mathrm{Spec}(R)\};$$

that is, $\dim(R)$ is the length of a longest prime chain in R.

If R is a Dedekind ring, then $\dim(R) = 1$, for every nonzero prime is a maximal ideal; if R is a domain, then $\dim(R) = 0$ if and only if R is a field. The next proposition characterizes the noetherian rings of Krull dimension 0.

Proposition 11.159. *If R Is a noetherian ring, then* $\dim(R) = 0$ *if and only if every finitely generated R-module M has a composition series.*

Proof. Assume that R is noetherian with Krull dimension 0. Since R is noetherian, Corollary 6.120(iii) says that there are only finitely many minimal prime ideals. Since $\dim(R) = 0$, every prime ideal is a minimal prime ideal (and a maximal ideal). We conclude that R has only finitely many prime ideals, say, $\mathfrak{p}_1, \ldots, \mathfrak{p}_n$. Now $\mathrm{nil}(R) = \bigcap_{i=1}^n \mathfrak{p}_i$ is nilpotent, by Exercise 11.39 on page 938; say, $(\mathrm{nil}(R))^m = \{0\}$. Define

$$N = \mathfrak{p}_1 \cdots \mathfrak{p}_n \subseteq \mathfrak{p}_1 \cap \cdots \cap \mathfrak{p}_n = \mathrm{nil}(R),$$

so that

$$N^m = (\mathfrak{p}_1 \cdots \mathfrak{p}_n)^m = \{0\}.$$

Let M be a finitely generated R-module, and consider the chain

$$M \supseteq \mathfrak{p}_1 M \supseteq \mathfrak{p}_1\mathfrak{p}_2 M \supseteq \cdots \supseteq NM.$$

The factor module $\mathfrak{p}_1 \cdots \mathfrak{p}_{i-1} M/\mathfrak{p}_1 \cdots \mathfrak{p}_i M$ is an $(R/\mathfrak{p}_i)$-module; that is, it is a vector space over the field $R/\mathfrak{p}_i$ (for $\mathfrak{p}_i$ is a maximal ideal). Since M is finitely generated, the factor module is finite-dimensional, and so the chain can be refined so that all the factor modules are simple. Finally, repeat this argument for the chains

$$N^j M \supseteq \mathfrak{p}_1 N^j M \supseteq \mathfrak{p}_1\mathfrak{p}_2 N^j M \supseteq \cdots \supseteq N^{j+1} M.$$

Since $N^m = \{0\}$, we have constructed a composition series for M.

Conversely, if every finitely generated R-module has a composition series, then the cyclic R-module R has a composition series; say, of length ℓ. It follows that any ascending chain of ideals has length at most ℓ, and so R is noetherian. To prove that $\dim(R) = 0$,

we must show that R does not contain any prime ideals $\mathfrak{p} \supsetneq \mathfrak{q}$. Passing to the quotient ring $R/\mathfrak{q}$, we may restate the hypotheses: R is a domain having a nonzero prime ideal as well as a composition series $R \supseteq I_1 \supseteq \cdots \supseteq I_d \neq \{0\}$. The last ideal I_d is a minimal ideal; choose a nonzero element $x \in I_d$. Of course, $xI_d \subseteq I_d$; since R is a domain, $xI_d \neq \{0\}$, so that minimality of I_d gives $xI_d = I_d$. Hence, there is $y \in I_d$ with $xy = x$; that is, $1 = y \in I_d$, and so $I_d = R$. We conclude that R is a field, contradicting its having a nonzero prime ideal. •

We are going to prove a theorem of W. Krull, the *principal ideal theorem*, which implies that every prime ideal in a noetherian ring has finite height. Our proof is Kaplansky's adaptation of a proof by D. Rees. We begin with a technical lemma.

Lemma 11.160. *Let a and b be nonzero elements in a domain R. If there exists $c \in R$ such that $ca^2 \in (b)$ implies $ca \in (b)$, then the series $(a, b) \supseteq (a) \supseteq (a^2)$ and $(a^2, b) \supseteq (a^2, ab) \supseteq (a^2)$ have isomorphic factor modules.*[7]

Proof. Now $(a, b)/(a) \cong (a^2, ab)/(a^2)$, for multiplication by a sends (a, b) onto (a^b, ab) and (a) onto (a^2).

The module $(a)/(a^2)$ is cyclic with annihilator (a); that is, $(a)/(a^2) \cong R/(a)$. The module $(a^2, b)/(a^2, ab)$ is also cyclic, for the generator a^2 lies in (a^2, ab). Now $A = \operatorname{ann}\big((a^2, b)/(a^2, ab)\big)$ contains (a), and so it suffices to prove that $A = (a)$; that is, if $cb = ua^2 + vab$, then $c \in (a)$. This equation gives $ua^2 \in (b)$, and so the hypothesis give $ua = rb$ for some $r \in R$. Substituting, $cb = rab + vab$, and canceling b gives $c = ra + va \in (a)$. Therefore, $(a)/(a^2) \cong (a^2, b)/(a^2, ab)$. •

Recall that a prime ideal $\mathfrak{p}$ is *minimal* over an ideal I if $I \subseteq \mathfrak{p}$ and there is no prime ideal $\mathfrak{q}$ with $I \subseteq \mathfrak{q} \subsetneq \mathfrak{p}$.

Theorem 11.161 (Principal Ideal Theorem). *Let (a) be a proper ideal in a noetherian ring R, and let $\mathfrak{p}$ be a prime ideal minimal over (a). Then* $\operatorname{ht}(\mathfrak{p}) \leq 1$.

Proof. If, on the contrary, $\operatorname{ht}(\mathfrak{p}) \geq 2$, then there is a prime chain

$$\mathfrak{p} \supsetneq \mathfrak{p}_1 \supsetneq \mathfrak{p}_2.$$

We normalize the problem in two ways. First, replace R by $R/\mathfrak{p}_2$; second, localize at $\mathfrak{p}/\mathfrak{p}_2$. We now modify the hypotheses accordingly: R is a local domain whose maximal ideal $\mathfrak{m}$ is minimal over a proper principal ideal (x), and there is a prime ideal $\mathfrak{q}$ with

$$\mathfrak{m} \supsetneq \mathfrak{q} \supsetneq (0).$$

Choose a nonzero element $b \in \mathfrak{q}$, and define

$$I_i = ((b) : x^i) = \{c \in R : cx^i \in (b)\}.$$

[7]Our notation for ideals is not consistent. The principal ideal generated by an element $a \in R$ is sometimes denoted by (a) and sometimes denoted by Ra.

There is an ascending chain $I_1 \subseteq I_2 \subseteq \cdots$, that must stop, because R is noetherian: say, $I_n = I_{n+1} = \cdots$. It follows that if $c \in I_{2n}$, then $c \in I_n$; that is, if $cx^{2n} \in (b)$, then $cx^n \in (b)$. If we set $a = x^n$, then $ca^2 \in (b)$ implies $ca \in (b)$.

If $R^* = R/(a^2)$, then $\dim(R^*) = 0$, for it has exactly one prime ideal. By Proposition 11.159, the R^*-module $(a, b)/(a^2)$ (as every finitely generated R^*-module) has finite length ℓ (the length of its composition series). But Lemma 11.160 implies that both (a, b) and its submodule (a^2, b) have length ℓ. The Jordan-Hölder theorem says this can happen only if $(a^2, b) = (a, b)$, which forces $a \in (a^2, b)$: there are $s, t \in R$ with $a = sa^2 + tb$. Since $sa \in \mathfrak{m}$, the element $1 - sa$ is a unit (for $(R, \mathfrak{m})$ is a local ring). Hence, $-a(1 - sa) = tb \in (b)$ gives $a \in (b) \subseteq \mathfrak{q}$. But $a = x^n$ gives $x \in \mathfrak{q}$, contradicting $\mathfrak{m}$ being a prime ideal minimal over (x). •

We now generalize the principal ideal theorem to finitely generated ideals.

Theorem 11.162 (Generalized Principal Ideal Theorem). *Let $I = (a_1, \ldots, a_n)$ be a proper ideal in a noetherian ring R, and let $\mathfrak{p}$ be a prime ideal minimal over I. Then* $\mathrm{ht}(\mathfrak{p}) \leq n$.

Proof. The hypotheses still hold after localizing at $\mathfrak{p}$, so we may assume that R is a local ring with $\mathfrak{p}$ as its maximal ideal.

The proof is by induction on $n \geq 1$, the base step being the principal ideal theorem. Let $I = (a_1, \ldots, a_{n+1})$, and assume, by way of contradiction, that $\mathrm{ht}(\mathfrak{p}) > n + 1$: there is a prime chain

$$\mathfrak{p} = \mathfrak{p}_0 \supsetneq \mathfrak{p}_1 \supsetneq \cdots \supsetneq \mathfrak{p}_{n+1}.$$

We may assume there are no prime ideals strictly between $\mathfrak{p}$ and $\mathfrak{p}_1$, for the module $\mathfrak{p}/\mathfrak{p}_1$ has ACC. Now $I \not\subseteq \mathfrak{p}_1$, because $\mathfrak{p}$ is a prime ideal minimal over I. Reindexing the generators of I if necessary, $a_1 \notin \mathfrak{p}_1$. Hence, $(a_1, \mathfrak{p}_1) \supsetneq \mathfrak{p}_1$. We claim that $\mathfrak{p}$ is the only prime ideal containing $(a_1, \mathfrak{p}_1)$; there can be no prime ideal $\mathfrak{p}'$ with $(a_1, \mathfrak{p}_1) \subseteq \mathfrak{p}' \subseteq \mathfrak{p}$ (the second inclusion holds because $\mathfrak{p}$ is the only maximal ideal in R) because there are no prime ideals strictly between $\mathfrak{p}$ and $\mathfrak{p}_1$. Therefore, in the ring $R/(a_1, \mathfrak{p}_1)$, the image of $\mathfrak{p}$ is the unique nonzero prime ideal. As such, it must be the nilradical, and hence it is nilpotent, by Exercise 11.39 on page 938. There is an integer m with $\mathfrak{p}^m \subseteq (a_1, \mathfrak{p}_1)$, and so there are equations

$$a_i^m = r_i a_1 + b_i, \qquad r_i \in R, b_i \in \mathfrak{p}_1, \text{and } i \geq 2. \tag{7}$$

Define $J = (b_2, \ldots, b_{n+1})$. Now $J \subseteq \mathfrak{p}_1$, while $\mathrm{ht}(\mathfrak{p}_1) > n$. By induction, $\mathfrak{p}_1$ cannot be a prime ideal minimal over J, and so there exists a prime ideal $\mathfrak{q}$ minimal over J:

$$J \subseteq \mathfrak{q} \subsetneq \mathfrak{p}_1.$$

Now $a_i^m \in (a_1, \mathfrak{q})$ for all i, by Eq. (7). Thus, any prime ideal $\mathfrak{p}'$ containing $(a_1, \mathfrak{q})$ must contain all a_i^m, hence all a_i, and hence I. As $\mathfrak{p}$ is the unique maximal ideal, $I \subseteq \mathfrak{p}' \subseteq \mathfrak{p}$. But $\mathfrak{p}$ is a prime ideal minimal over I, and so $\mathfrak{p}' = \mathfrak{p}$. Therefore, $\mathfrak{p}$ is the unique prime ideal containing $(a_1, \mathfrak{q})$. If $R^* = R/\mathfrak{q}$, then $\mathfrak{p}^* = \mathfrak{p}/\mathfrak{q}$ is a prime ideal minimal over the principal ideal $(a_1 + \mathfrak{q})$. On the other hand, $\mathrm{ht}(\mathfrak{p}^*) \geq 2$, for $\mathfrak{p}^* \supsetneq \mathfrak{p}_1^* \supsetneq \{0\}$ is a prime chain, where $\mathfrak{p}_1^* = \mathfrak{p}_1/\mathfrak{q}$. This contradiction to the principal ideal theorem completes the proof. •

Corollary 11.163. *If R is a noetherian ring, then every prime ideal has finite height, and so* $\operatorname{Spec}(R)$ *has* DCC.

Proof. Every prime ideal $\mathfrak{p}$ is finitely generated, because R is noetherian; say, $\mathfrak{p} = (a_1, \ldots, a_n)$. But $\mathfrak{p}$ is a minimal prime ideal over itself, so that Theorem 11.162 gives $\operatorname{ht}(\mathfrak{p}) \le n$. •

A noetherian ring may have infinite Krull dimension, for there may be no uniform bound on the length of prime chains. We will see that this cannot happen for local rings.

The generalized principal ideal theorem bounds the height of a prime ideal that is minimal over an ideal; the next result bounds the height of a prime ideal that merely contains an ideal.

Corollary 11.164. *Let R be a noetherian ring, let $I = (a_1, \ldots, a_n)$ be an ideal in R, and let $\mathfrak{p}$ be a prime ideal in R containing I. If* $\operatorname{ht}(\mathfrak{p}/I)$ *denotes the height of $\mathfrak{p}/I$ in R/I, then*

$$\operatorname{ht}(\mathfrak{p}) \le n + \operatorname{ht}(\mathfrak{p}/I).$$

Proof. The proof is by induction on $h = \operatorname{ht}(\mathfrak{p}/I) \ge 0$. If $h = 0$, then Exercise 11.76 on page 1011 says that $\mathfrak{p}$ is minimal over I, and so the base step is the generalized principal ideal theorem. For the inductive step $h > 0$, $\mathfrak{p}$ is not minimal over I. By Corollary 6.120(iii), there are only finitely many minimal primes in R/I, and so Exercise 11.76 says that there are only finitely many prime ideals minimal over I; say, $\mathfrak{q}_1, \ldots, \mathfrak{q}_s$. Since $\mathfrak{p}$ is not minimal over I, $\mathfrak{p} \not\subseteq \mathfrak{q}_i$ for any i; hence, Proposition 6.14 says that $\mathfrak{p} \not\subseteq \mathfrak{q}_1 \cup \cdots \cup \mathfrak{q}_s$, and so there is $y \in \mathfrak{p}$ with $y \notin \mathfrak{q}_i$ for any i. Define $J = (I, y)$.

We now show, in R/J, that $\operatorname{ht}(\mathfrak{p}/J) \le h - 1$. Let

$$\mathfrak{p}/J \supsetneq \mathfrak{p}_1/J \supsetneq \cdots \supsetneq \mathfrak{p}_r/J$$

be a prime chain in R/J. Since $I \subsetneq J$, there is a surjective ring map $R/I \to R/J$. The prime chain lifts to a prime chain in R/I:

$$\mathfrak{p}/I \supsetneq \mathfrak{p}_1/I \supsetneq \cdots \supsetneq \mathfrak{p}_r/I.$$

Now $\mathfrak{p}_r \supseteq J \supsetneq I$, and $J = (I, y)$ does not contain any $\mathfrak{q}_i$. But the ideals $\mathfrak{q}_i/I$ are the minimal prime ideals in R/I, by Exercise 11.76, so that $\mathfrak{p}_r$ is not a minimal prime ideal in R. Therefore, there is a prime chain starting at $\mathfrak{p}$ of length $r + 1$. We conclude that $r + 1 \le h$, and so $\operatorname{ht}(\mathfrak{p}/J) \le h - 1$.

Since $J = (I, y) = (a_1, \ldots, a_n, y)$ is generated by $n + 1$ elements, the inductive hypothesis gives

$$\begin{aligned} \operatorname{ht}(\mathfrak{p}) &\le n + 1 + \operatorname{ht}(\mathfrak{p}/J) \\ &= (n + 1) + (h - 1) = n + h = n + \operatorname{ht}(\mathfrak{p}/I). \quad \bullet \end{aligned}$$

When we say, in the next proposition, that a generating set X of an ideal I is *minimal*, we mean that no proper subset of X generates I.

If $(R, \mathfrak{m}, k)$ is a local ring, then $\mathfrak{m}/\mathfrak{m}^2$ is an $(R/\mathfrak{m})$-module; that is, it is a vector space over k.

Proposition 11.165. *Let $(R, \mathfrak{m}, k)$ be a noetherian local ring.*

(i) *Elements $x_1, \ldots, x_d$ form a minimal generating set for $\mathfrak{m}$ if and only if the cosets $x_i^* = x_i + \mathfrak{m}^2$ form a basis of $\mathfrak{m}/\mathfrak{m}^2$.*

(ii) *Any two minimal generating sets of $\mathfrak{m}$ have the same number of elements.*

Proof. (i) If $x_1, \ldots, x_d$ is a minimal generating set for $\mathfrak{m}$, then $X^* = x_1^*, \ldots, x_d^*$ spans the vector space $\mathfrak{m}/\mathfrak{m}^2$. If X^* is linearly dependent, then there is some $x_i^* = \sum_{j \neq i} r_j' x_J^*$, where $r_j' \in k$. Lifting this equation to $\mathfrak{m}$, we have $x_i \in \sum_{j \neq i} r_j x_j + \mathfrak{m}^2$. Thus, if $B = \langle x_j : j \neq i \rangle$, then $B + \mathfrak{m}^2 = \mathfrak{m}$. Hence,

$$\mathfrak{m}\,(\mathfrak{m}/B) = (B + \mathfrak{m}^2)/B = \mathfrak{m}/B.$$

By Nakayama's lemma, $\mathfrak{m}/B = \{0\}$, and so $\mathfrak{m} = B$. This contradicts $x_1, \ldots, x_d$ being a minimal generating set. Therefore, X^* is linearly independent, and hence it is a basis of $\mathfrak{m}/\mathfrak{m}^2$.

Conversely, assume that $x_1^*, \ldots, x_d^*$ is a basis of $\mathfrak{m}/\mathfrak{m}^2$, where $x_i^* = x_i + \mathfrak{m}^2$. If we define $A = \langle x_1, \ldots, x_d \rangle$, then $A \subseteq \mathfrak{m}$. If $y \in \mathfrak{m}$, then $y^* = \sum r_i' x_i^*$, where $r_i' \in k$, so that $y \in A + \mathfrak{m}^2$. Hence, $\mathfrak{m} = A + \mathfrak{m}^2$, and, as in the previous paragraph, Nakayama's lemma gives $\mathfrak{m} = A$; that is, $x_1, \ldots, x_d$ generate $\mathfrak{m}$. If a proper subset of $x_1, \ldots, x_d$ generates $\mathfrak{m}$, then the vector space $\mathfrak{m}/\mathfrak{m}^2$ could be generated by fewer than d elements, contradicting $\dim_k(\mathfrak{m}/\mathfrak{m}^2) = d$.

(ii) The number of elements in any minimal generating set is $\dim_k(\mathfrak{m}/\mathfrak{m}^2)$. •

Definition. If $(R, \mathfrak{m}, k)$ is a noetherian local ring, then $\mathfrak{m}/\mathfrak{m}^2$ is a finite-dimensional vector space over k. Write

$$\mu(\mathfrak{m}) = \dim_k(\mathfrak{m}/\mathfrak{m}^2).$$

Proposition 11.165 shows that all minimal generating sets of $\mathfrak{m}$ have the same number of elements, namely, $\mu(\mathfrak{m})$.

Corollary 11.166. *If $(R, \mathfrak{m})$ is a noetherian local ring, then* $\mathrm{ht}(\mathfrak{m}) \leq \mu(\mathfrak{m})$, *and*

$$\dim(R) \leq \mu(\mathfrak{m}).$$

Proof. If $\mu(\mathfrak{m}) = d$, then $\mathfrak{m} = (x_1, \ldots, x_d)$. Since $\mathfrak{m}$ is obviously a minimal prime over itself, Theorem 11.162, the generalized principal ideal theorem, gives $\mathrm{ht}(\mathfrak{m}) \leq d = \mu(\mathfrak{m})$.

If $\mathfrak{p} \neq \mathfrak{m}$ is a prime ideal in R, then any prime chain, $\mathfrak{p} = \mathfrak{p}_0 \supsetneq \mathfrak{p}_1 \supsetneq \cdots \supsetneq \mathfrak{p}_h$, can be lengthened by to a prime chain $\mathfrak{m} \supsetneq \mathfrak{p}_0 \supsetneq \mathfrak{p}_1 \supsetneq \cdots \supsetneq \mathfrak{p}_h$ of length $h + 1$. Therefore, $h < \mu(\mathfrak{m})$, and so $\dim(R) = \mathrm{ht}(\mathfrak{m}) \leq \mu(\mathfrak{m})$. •

Definition. A ***regular local ring*** is a noetherian local ring $(R, \mathfrak{m})$ such that

$$\dim(R) = \mu(\mathfrak{m}).$$

It is clear that every field is a regular local ring of dimension 0, and every DVR is a regular local ring of dimension 1. It is not clear from the definition whether there are any other examples. The coming notion of *regular sequence* will enable us to better understand regular local rings. Recall that if M is an R-module, then an element $c \in R$ is called *regular* on M if the map $M \to M$, given by $m \mapsto cm$, is an injection; that is, $cm = 0$ implies $m = 0$.

Definition. A sequence $x_1, \ldots, x_n$ in a commutative ring R is an ***M-regular sequence*** if x_1 is regular on M, x_2 is regular on $M/(x_1)M$, x_3 is regular on $M/(x_1, x_2)M$,..., x_n is regular on $M/(x_1, \cdots, x_{n-1})M$. If $M = R$, then $x_1, \ldots, x_n$ is also called an ***R-sequence***.

For example, if $R = k[x_1, \ldots, x_n]$ is a polynomial ring over a field k, then it is easy to see that $x_1, \ldots, x_n$ is an R-sequence.

Exercise 11.75 on page 1011 gives an example of a permutation of an R-sequence that is not an R-sequence. However, if R is local, then every permutation of an R-sequence is also an R-sequence (see Bruns–Herzog, *Cohen–Macaulay Rings*, page 5).

The generalized principal ideal theorem gives an upper bound on the height of a prime ideal; the next lemma gives a lower bound.

Lemma 11.167. *Let R be a commutative ring.*

(i) *If $x \in R$ is not a zero divisor, then x lies in no minimal prime ideal.*

(ii) *If $\mathfrak{p}$ is a prime ideal in R and $x \in \mathfrak{p}$ is not a zero divisor, then*

$$\operatorname{ht}(\mathfrak{p}/(x)) + 1 \leq \operatorname{ht}(\mathfrak{p}).$$

(iii) *If a prime ideal $\mathfrak{p}$ in R contains an R-sequence $x_1, \ldots, x_d$, then*

$$d \leq \operatorname{ht}(\mathfrak{p}).$$

Proof. (i) Suppose, on the contrary, that $\mathfrak{p}$ is a nonzero minimal prime ideal containing x. Now $R_\mathfrak{p}$ is a ring with only one nonzero prime ideal, namely, $\mathfrak{p}_\mathfrak{p}$, which must be the nilradical. Thus, $x/1$, as every element in $\mathfrak{p}_\mathfrak{p}$, is nilpotent. If $x^m/1 = 0$ in $R_\mathfrak{p}$, then there is $\sigma \notin \mathfrak{p}$ (so that $\sigma \neq 0$) with $\sigma x = 0$, contradicting x not being a zero divisor.

(ii) If $h = \operatorname{ht}(\mathfrak{p}/(x))$, then there is a prime chain in $R/(x)$:

$$\mathfrak{p}/(x) \supsetneq \mathfrak{p}_1/(x) \supsetneq \cdots \supsetneq \mathfrak{p}_h/(x).$$

Lifting back to R, there is a prime chain $\mathfrak{p} \supsetneq \mathfrak{p}_1 \supsetneq \cdots \supsetneq \mathfrak{p}_h$ with $\mathfrak{p}_h \supseteq (x)$. Since x is not a zero divisor, part (i) says that $\mathfrak{p}_h$ is not a minimal prime. Therefore, there exists a prime ideal $\mathfrak{p}_{h+1}$ properly contained in $\mathfrak{p}_h$, which shows that $\operatorname{ht}(\mathfrak{p}) \geq h + 1$.

(iii) The proof is by induction on $d \geq 1$. For the base step $d = 1$, suppose, on the contrary, that $\operatorname{ht}(\mathfrak{p}) = 0$; then $\mathfrak{p}$ is a minimal prime ideal, and this contradicts part (i). For the inductive step, part (ii) gives $\operatorname{ht}(\mathfrak{p}/(x_1)) + 1 \leq \operatorname{ht}(\mathfrak{p})$. Now $\mathfrak{p}/(x_1)$ contains an $(R/(x_1))$-sequence $x_2 + (x_1), \ldots, x_d + (x_1)$, by Exercise 11.79(ii) on page 1012, so that the inductive hypothesis gives $d - 1 \leq \operatorname{ht}(\mathfrak{p}/(x_1))$. Therefore, part (ii) gives $d \leq \operatorname{ht}(\mathfrak{p})$. •

Proposition 11.168. *Let $(R, \mathfrak{m})$ be a noetherian local ring. If $\mathfrak{m}$ can be generated by an R-sequence $x_1, \ldots, x_d$, then R is a regular local ring and*

$$d = \dim(R) = \mu(\mathfrak{m}).$$

Remark. We will soon prove the converse: In a regular local ring, the maximal ideal can be generated by an R-sequence. ◂

Proof. Consider the inequalities

$$d \leq \operatorname{ht}(\mathfrak{m}) \leq \mu(\mathfrak{m}) \leq d.$$

The first inequality holds by Lemma 11.167; the second by Corollary 11.166, and the third by Proposition 11.165. It follows that all the inequalities are, in fact, equalities, and the proposition follows because $\dim(R) = \operatorname{ht}(\mathfrak{m})$. •

Example 11.169.

Let k be a field, and let $R = k[[x_1, \ldots, x_r]]$ be the ring of formal power series in r variables $x_1, \ldots, x_r$. Recall that an element $f \in R$ is a sequence

$$f = (f_0, f_1, f_2, \ldots, f_n, \ldots),$$

where f_n is a homogeneous polynomial of total degree n in $k[x_1, \ldots, x_r]$, and that multiplication is defined by

$$(f_0, f_1, f_2, \ldots)(g_0, g_1, g_2, \ldots) = (h_0, h_1, h_2, \ldots),$$

where $h_n = \sum_{i+j=n} f_i g_j$. We claim that R is a local ring with maximal ideal $\mathfrak{m} = (x_1, \ldots, x_r)$ and residue field k. First, $R/\mathfrak{m} \cong k$, so that $\mathfrak{m}$ is a maximal ideal. To see that $\mathfrak{m}$ is the unique maximal ideal, it suffices to prove that if $f \in R$ and $f \notin \mathfrak{m}$, then f is a unit. Now $f \notin \mathfrak{m}$ if and only if $f_0 \neq 0$, and we now show that f is a unit if and only if $f_0 \neq 0$. If $fg = 1$, then $f_0 g_0 = 1$, and $f_0 \neq 0$; conversely, if $f_0 \neq 0$, we can solve $(f_0, f_1, f_2, \ldots)(g_0, g_1, g_2, \ldots) = 1$ recursively for g_n, and $fg = 1$ if we define $g = (g_0, g_1, g_2, \ldots)$.

Exercise 11.83 on page 1012 shows that the ring R is noetherian. But $R/(x_1, \ldots, x_{i-1})$ is a domain, because it is isomorphic to $k[[x_i, \ldots, x_r]]$, and so x_i is a regular element on it. Hence, Proposition 11.168 shows that $R = k[[x_1, \ldots, x_r]]$ is a regular local ring, for $x_1, \ldots, x_r$ is an R-sequence. ◂

The next lemmas prepare us for induction.

Lemma 11.170. *Let $(R, \mathfrak{m}, k)$ be a noetherian local ring, and let $x \in \mathfrak{m} - \mathfrak{m}^2$.*

(i) *If $x_1 + (x), \ldots, x_s + (x)$ is a minimal generating set of $\mathfrak{m}/x\mathfrak{m}$, then $x, x_1, \ldots, x_s$ is a minimal generating set of $\mathfrak{m}$.*

(ii)
$$\mu(\mathfrak{m}/x\mathfrak{m}) = \mu(\mathfrak{m}) - 1.$$

Proof. (i) Write $\overline{R} = R/(x)$, $\overline{\mathfrak{m}} = \mathfrak{m}/x\mathfrak{m}$, and $\overline{r} = r + (x)$ for all $r \in R$. To see that $x, x_1, \ldots, x_s$ generate $\mathfrak{m}$, let $y \in \mathfrak{m}$. Then $\overline{y} = \sum_i r_i' \overline{x}_i$, where $r_i' \in k$. Lifting to R gives $y - \sum_i r_i x_i \in (x)$, where $r_i' = r_i + \mathfrak{m}$. Therefore, there is $r \in R$ with $y = rx + \sum r_i x_i$.

To prove minimality, Proposition 11.165 says that it suffices to show that the cosets $x^* = x + \mathfrak{m}^2, x_i^* = x_i + \mathfrak{m}^2$ form a basis of $\mathfrak{m}/\mathfrak{m}^2$. These elements span $\mathfrak{m}/\mathfrak{m}^2$ because $x, x_1, \ldots, x_s$ generate $\mathfrak{m}$. To prove linear independence, assume that $a'x^* + \sum a_i' x_i^* = 0$, where $a', a_i' \in k$. Lifting to R, we have

$$ax + \sum a_i x_i \in \mathfrak{m}^2, \tag{8}$$

and we must show that $a, a_i \in \mathfrak{m}$ (for then $a', a_i' = 0$ in $k = R/\mathfrak{m}$). In $\overline{R} = R/(x)$, this relation becomes
$$\sum_i \overline{a}_i \overline{x}_i \in \overline{\mathfrak{m}}^2.$$

As $\overline{x}_1, \ldots, \overline{x}_s$ is a basis of $\overline{\mathfrak{m}}/\overline{\mathfrak{m}}^2$, we have $a_i' = 0$ for all i; that is, $a_i \in \mathfrak{m}$ for all i. It follows from Eq. (8) that $ax \in \mathfrak{m}^2$. Since $x \notin \mathfrak{m}^2$, it follows that $a \in \mathfrak{m}$, as desired.

(ii) This follows at once from part (i). •

Lemma 11.171. *Let $(R, \mathfrak{m})$ be a regular local ring. If $x \in \mathfrak{m} - \mathfrak{m}^2$, then $R/(x)$ is regular and* $\dim(R/(x)) = \dim(R) - 1$.

Proof. Since R is regular, we have $\dim(R) = \mu(\mathfrak{m})$. Let us note at the outset that $\dim(R) = \mathrm{ht}(\mathfrak{m})$. We must show that $\mathrm{ht}(\mathfrak{m}^*) = \mu(\mathfrak{m}^*)$, where $\mathfrak{m}^* = \mathfrak{m}/(x)$. By Corollary 11.164, $\mathrm{ht}(\mathfrak{m}) \le \mathrm{ht}(\mathfrak{m}^*) + 1$. Hence,

$$\begin{aligned} \mathrm{ht}(\mathfrak{m}) - 1 &\le \mathrm{ht}(\mathfrak{m}^*) \\ &\le \mu(\mathfrak{m}^*) \\ &= \mu(\mathfrak{m}) - 1 \\ &= \mathrm{ht}(\mathfrak{m}) - 1. \end{aligned}$$

The next to last equation is Lemma 11.170; the last equation holds because R is regular. Therefore, $\dim(R^*) = \mathrm{ht}(\mathfrak{m}^*) = \mu(\mathfrak{m}^*)$, and so $R^* = R/(x)$ is regular with $\dim(R/(x)) = \dim(R) - 1$. •

We are now going to prove that regular local rings are domains, and we will then use this to prove the converse of Proposition 11.168.

Proposition 11.172. *Every regular local ring $(R, \mathfrak{m})$ is a domain.*

Proof. The proof is by induction on $d = \dim(R)$. If $d = 0$, then R is a field, by Exercise 11.78 on page 1012. If $d > 0$, let $\mathfrak{p}_1, \ldots, \mathfrak{p}_s$ be the minimal prime ideals in R (there are only finitely many such, by Corollary 6.120). If $\mathfrak{m} - \mathfrak{m}^2 \subseteq \mathfrak{p}_1 \cup \cdots \cup \mathfrak{p}_s$, then Proposition 6.14 would give $\mathfrak{m} \subseteq \mathfrak{p}_i$, which cannot occur because $d = \mathrm{ht}(\mathfrak{m}) > 0$. Therefore, there is $x \in \mathfrak{m} - \mathfrak{m}^2$ with $x \notin \mathfrak{p}_i$ for all i. By Lemma 11.171, $R/(x)$ is regular of dimension $d - 1$. The inductive hypothesis gives $R/(x)$ a domain, and so (x) is a prime ideal. It follows that (x) contains a minimal prime ideal; say, $\mathfrak{p}_i \subseteq (x)$.

If $\mathfrak{p}_i = \{0\}$, then $\{0\}$ is a prime ideal and R is a domain. Hence, we may assume that $\mathfrak{p}_i \neq \{0\}$. For each nonzero $y \in \mathfrak{p}_i$, there exists $r \in R$ with $y = rx$. Since $x \notin \mathfrak{p}_i$, we have $r \in \mathfrak{p}_i$, so that $y \in x\mathfrak{p}_i$. Thus, $\mathfrak{p}_i \subseteq x\mathfrak{p}_i \subseteq \mathfrak{m}\mathfrak{p}_i$. As the reverse inclusion $\mathfrak{m}\mathfrak{p}_i \subseteq \mathfrak{p}_i$ is always true, we have $\mathfrak{p}_i = \mathfrak{m}\mathfrak{p}_i$. Nakayama's lemma now applies, giving $\mathfrak{p}_i = \{0\}$, a contradiction. •

Proposition 11.173. *A noetherian local ring $(R, \mathfrak{m}, k)$ is regular if and only if $\mathfrak{m}$ is generated by an R-sequence $x_1, \ldots, x_d$. Moreover, in this case,*

$$d = \mu(\mathfrak{m}).$$

Proof. We have already proven sufficiency, in Proposition 11.168. If R is regular, we prove the result by induction on $d \geq 1$, where $d = \dim(R)$. The base step holds, for R is a domain and so x is a regular element; that is, x is not a zero divisor. For the inductive step, the ring $R/(x)$ is regular of dimension $d - 1$, by Lemma 11.171. Therefore, its maximal ideal is generated by an $(R/(x))$-sequence $x_1^*, \ldots, x_{d-1}^*$. By Lemma 11.170, a minimal generating set for $\mathfrak{m}$ is $x, x_1, \ldots, x_{d-1}$. Finally, this is an R-sequence, by Exercise 11.79(i) on page 1012, because x is not a zero divisor. •

We are now going to characterize regular local rings by their global dimension.

Lemma 11.174. *Let $(R, \mathfrak{m}, k)$ be a local ring. If A is an R-module with $pd(A) = n$ and if $x \in \mathfrak{m}$ is A-regular, then $pd(A/xA) = n + 1$.*

Proof. Since x is A-regular, there is an exact sequence

$$0 \to A \xrightarrow{\mu_x} A \to A/xA \to 0,$$

where $\mu_x \colon a \mapsto xa$. By Lemma 11.147, we have $pd(A/xA) \leq n + 1$.

Consider the portion of the long exact sequence arising from tensoring by k:

$$0 = \mathrm{Tor}_{n+1}^R(A, k) \to \mathrm{Tor}_{n+1}^R(A/xA, k) \xrightarrow{\partial} \mathrm{Tor}_n^R(A, k) \xrightarrow{(\mu_x)_*} \mathrm{Tor}_n^R(A, k).$$

Now $pd(A) \leq n$ if and only if $\mathrm{Tor}_{n+1}^R(A, k) = \{0\}$, by Lemma 11.155, and so the first term is $\{0\}$. The induced map $(\mu_x)_*$ is multiplication by x. However, if $\mu_x' \colon k \to k$ is multiplication by x, then $x \in \mathfrak{m}$ implies $\mu_x' = 0$; therefore, $(\mu_x)_* = (\mu_x')_* = 0$. Exactness now gives $\partial \colon \mathrm{Tor}_{n+1}^R(A/xA, k) \to \mathrm{Tor}_n^R(A, k)$ an isomorphism. Since $pd(A) = n$, we have $\mathrm{Tor}_n^R(A, k) \neq \{0\}$, so that $\mathrm{Tor}_{n+1}^R(A/xA, k) \neq \{0\}$. Therefore, $pd(A/xA) \geq n + 1$, as desired. •

Proposition 11.175. *If $(R, \mathfrak{m}, k)$ is a regular local ring, then*

$$D(R) = \mu(\mathfrak{m}) = \dim(R).$$

Proof. Since R is regular, Proposition 11.173 says that $\mathfrak{m}$ can be generated by an R-sequence $x_1, \ldots, x_d$. Applying Lemma 11.174 to the modules R, $R/(x_1)$, $R/(x_1, x_2), \ldots$, $R/(x_1, \ldots, x_d) = R/\mathfrak{m} = k$, we see that $pd(k) = d$. By Proposition 11.168, $d = \mu(\mathfrak{m}) = \dim(R)$. On the other hand, Theorem 11.157(ii) gives $d = pd(k) = D(R)$. •

The converse of Proposition 11.175, A noetherian local ring of finite global dimension is regular, is more difficult to prove.

Lemma 11.176. *Let $(R, \mathfrak{m}, k)$ be a noetherian local ring of finite global dimension. If $\mu(\mathfrak{m}) \le D(R)$ and $D(R) \le d$, where d is the length of a longest R-sequence in $\mathfrak{m}$, then R is a regular local ring.*

Proof. By Corollary 11.166, $\dim(R) \le \mu(\mathfrak{m})$. By hypothesis, $\mu(\mathfrak{m}) \le D(R) \le d$, while Lemma 11.167 gives $d \le \mathrm{ht}(\mathfrak{m}) = \dim(R)$. Therefore, $\dim(R) = \mu(\mathfrak{m})$, and so R is a regular local ring. •

Let R be a noetherian ring, let M be a finitely generated R-module, and let I be an ideal such that $IM \ne M$. By Exercise 11.82 on page 1012, I contains a longest M-sequence (such sequences are usually called ***maximal M-sequences*** in I). We are going to prove, given an ideal I and a finitely generated R-module M, that all maximal M-sequences in I have the same length.

Definition. If R is a commutative ring, then an ***associated prime ideal*** of a nonzero R-module B is a prime ideal of the form $\mathrm{ann}(b)$ for some nonzero $b \in B$.

Lemma 11.177. *Let B be a nonzero finitely generated module over a noetherian ring R.*

(i) *The maximal elements in $\mathcal{F}(B) = \{\mathrm{ann}(b) : b \in B \text{ and } b \ne 0\}$ are associated prime ideals of B.*

(ii) *There are finitely many associated prime ideals of B, say, $\mathfrak{p}_1, \ldots, \mathfrak{p}_s$, such that*

$$Z(B) = \mathfrak{p}_1 \cup \cdots \cup \mathfrak{p}_s,$$

where $Z(B) = \{r \in R : rb = 0 \text{ for some nonzero } b \in B\}$.

Proof. (i) The set of ideals $\mathcal{F}(B)$ has maximal elements, because R is noetherian. Let $\mathrm{ann}(b)$ be such a maximal element. Suppose that $rs \in \mathrm{ann}(b)$, where $r, s \in R$ and $r \notin \mathrm{ann}(b)$. Now $\mathrm{ann}(b) \subseteq \mathrm{ann}(rb)$, for if $ub = 0$, then $u(rb) = 0$; by maximality, $\mathrm{ann}(b) = \mathrm{ann}(rb)$. Hence, $s \in \mathrm{ann}(rb)$ implies $s \in \mathrm{ann}(b)$, and so $\mathrm{ann}(b)$ is a prime ideal.

(ii) For each $r \in Z(B)$, there is a nonzero $b \in B$ with $rb = 0$; that is, $Z(B) = \bigcup_{\mathrm{ann}(b) \in \mathcal{F}(B)} \mathrm{ann}(b)$. If we denote the set of maximal elements in $\mathcal{F}(B)$ by $\mathfrak{M}$, then $Z(B) = \bigcup_{\mathfrak{p} \in \mathfrak{M}} \mathfrak{p}$, for every $\mathrm{ann}(b) \in \mathcal{F}(B)$ is contained in a maximal element.

It suffices to prove that $\mathfrak{M}$ is finite. Define $B' = \langle b : \text{ann}(b) \in \mathfrak{M}\rangle$. Now B' is finitely generated, for R noetherian implies that every submodule of a finitely generated R-module is itself finitely generated; let $B' = \langle b_1, \ldots, b_n\rangle$, and denote $\text{ann}(b_i)$ by $\mathfrak{p}_i$. Suppose there is $\mathfrak{q} = \text{ann}(b_0) \in \mathfrak{M}$ with $b_0 \neq b_i$ for $i = 1, \ldots, n$. As $b_0 \in B'$, there are $r_i \in R$ with $b_0 = \sum_i r_i b_i$. It follows that if $r \in \bigcap_i \mathfrak{p}_i$, then $rb_0 = 0$; that is, $\bigcap_i \mathfrak{p}_i \subseteq \text{ann}(b_0) = \mathfrak{q}$. Since $\mathfrak{q}$ is a prime ideal, Proposition 6.13 gives $\mathfrak{p}_i \subseteq \mathfrak{q}$ for some i. As $\mathfrak{p}_i$ is a maximal element in $\mathcal{F}(B)$, we have $\mathfrak{q} = \mathfrak{p}_i$, as desired. •

Remark. The set $\text{Ass}(B)$ of all associated primes of an R-module B is important in deeper studies [$\mathfrak{M}$ may be a proper subset of $\text{Ass}(B)$]. For example, it is related to primary decompositions (see Matsumura, *Commutative Ring Theory*, pages 39 - 42). ◂

The next lemma is a generalization of the observation that $\text{Hom}_{\mathbb{Z}}(\mathbb{I}_m, \mathbb{I}_n) = \{0\}$ if $(m, n) = 1$.

Lemma 11.178. *Let R be a commutative ring, and let A and B be R-modules.*

(i) *If* $\text{ann}(A)$ *contains a B-regular element, then* $\text{Hom}_R(A, B) = \{0\}$.

(ii) *Conversely, let R be noetherian, and let A and B be finitely generated R-modules. If* $\text{Hom}_R(A, B) = \{0\}$, *then* $\text{ann}(A)$ *contains a B-regular element.*

Proof. (i) If $r \in \text{ann}(A)$, then $ra = 0$ for all $a \in A$. Hence, for all $f \in \text{Hom}_R(A, B)$, we have $0 = f(ra) = rf(a)$. On the other hand, if r is B-regular, then $rf(a) = 0$ implies $f(a) = 0$, and so $f = 0$.

(ii) Assume, on the contrary, that $\text{ann}(A)$ contains no B-regular elements; that is, $\text{ann}(A) \subseteq Z(B)$. By Lemma 11.177, there are finitely many associated prime ideals of B, say, $\mathfrak{p}_1, \ldots, \mathfrak{p}_s$, such that $\text{ann}(A) \subseteq Z(B) = \mathfrak{p}_1 \cup \cdots \cup \mathfrak{p}_s$, and so Proposition 6.14 says that there is some $\mathfrak{p} = \mathfrak{p}_i$ with $\text{ann}(A) \subseteq \mathfrak{p}$.

Suppose that $A_\mathfrak{p} = \{0\}$. If $A = \langle a_1, \ldots, a_n\rangle$, then there are $\sigma_i \notin \mathfrak{p}$ with $\sigma_i a_i = 0$, by Proposition 11.25. Since $\mathfrak{p}$ is prime, $\sigma = \sigma_1\sigma_2 \cdots \sigma_n \notin \mathfrak{p}$. But $\sigma \in \text{ann}\, A = I \subseteq \mathfrak{p}$, and this is a contradiction. Therefore, $A_\mathfrak{p} \neq \{0\}$.

We wish to prove that $\text{Hom}_R(A, B) \neq \{0\}$. By Lemma 11.32, it suffices to prove that $\text{Hom}_R(A, B)_\mathfrak{p} \cong \text{Hom}_{R_\mathfrak{p}}(A_\mathfrak{p}, B_\mathfrak{p}) \neq \{0\}$. Thus, we may assume that $(R, \mathfrak{p}, k)$ is a local ring with maximal ideal $\mathfrak{p}$ and residue field k. Now there is an element $b \in B$ with $\text{ann}(b) = \mathfrak{p}$, so that $\langle b\rangle \cong R/\mathfrak{p} = k$. Hence, there is a nonzero map $\varphi \colon k \to B$ (taking $1 \mapsto b$). Since $A_\mathfrak{p} \neq \{0\}$, Nakayama's lemma gives $A/\mathfrak{p}A \neq \{0\}$. But $A/\mathfrak{p}A$ is a nonzero vector space over k, so there exists a nonzero map $A/\mathfrak{p}A \to k$. The composite of this map followed by φ is a nonzero map $A \to B$, and so $\text{Hom}_R(A, B) \neq \{0\}$. •

Lemma 11.179. *Let R be a commutative ring, let A and B be R-modules, and let $x_1, \ldots, x_n$ be a B-sequence in* $\text{ann}(A)$. *If $I = (x_1, \ldots, x_n)$, then*

$$\text{Hom}_R(A, B/IB) \cong \text{Ext}_R^n(A, B).$$

Proof. The proof is by induction on $n \geq 0$. We define $I = \{0\}$ in case $n = 0$, and so the base step holds. Assume now that $x_1, \ldots, x_{n+1}$ is a B-sequence in $\text{ann}(A)$, that $I = (x_1, \ldots, x_{n+1})$, and that $J = (x_1, \ldots, x_n)$. Observe first that there is an exact sequence $0 \to B \to B \to B/x_1B \to 0$, for x_1 is a regular element on B. Consider the portion of the long exact sequence, where x_1 is multiplication by x_1:

$$\text{Ext}^n_R(A, B) \xrightarrow{x_{1*}} \text{Ext}^n_R(A, B) \xrightarrow{\partial} \text{Ext}^n_R(A, B/x_1B) \to \text{Ext}^{n+1}_R(A, B) \xrightarrow{x_{1*}} \text{Ext}^{n+1}_R(A, B).$$

Since $x_1 \in \text{ann}(A)$, the induced map x_{1*} is the zero map, and there is a short exact sequence

$$0 \to \text{Ext}^n_R(A, B) \to \text{Ext}^n_R(A, B/x_1B) \xrightarrow{\partial} \text{Ext}^{n+1}_R(A, B) \to 0.$$

By induction, $\text{Hom}_R(A, B/JB) \cong \text{Ext}^n_R(A, B)$. Multiplication by $x_{n+1} \colon B/JB \to B/JB$ is an injection, because x_{n+1} is (B/JB)-regular, and left exactness of $\text{Hom}_R(A, \ )$ shows that $(x_{n+1})_*$ is an injection $\text{Hom}_R(A, B/JB) \to \text{Hom}_R(A, B/JB)$. On the other hand, $(x_{n+1})_*$ is the zero map, for $x_{n+1} \in \text{ann}(A)$. Hence, $\text{Hom}_R(A, B/JB) = \{0\}$, and $\text{Ext}^n_R(A, B) = \{0\}$. Therefore, $\partial \colon \text{Ext}^n_R(A, B/x_1B) \to \text{Ext}^{n+1}_R(A, B)$ is an isomorphism. By induction, if $B' = B/x_1B$, then $\text{Hom}_R(A, B'/(x_2, \ldots, x_{n+1})B') \cong \text{Ext}^n_R(A, B/x_1B)$. But

$$B'/(x_2, \ldots, x_{n+1})B' \cong (B/x_1B)/(IB/x_1B) \cong B/IB,$$

so that $\text{Hom}_R(A, B/IB) \cong \text{Ext}^n_R(A, B/x_1B)$. We conclude that $\text{Hom}_R(A, B/IB) \cong \text{Ext}^{n+1}_R(A, B)$, as desired. •

The following result is due to D. Rees.

Proposition 11.180. *Let R be a commutative noetherian ring, B a finitely generated R-module, and I an ideal with $IB \neq B$. Then all maximal B-sequences in I have the same length, say, g, where*

$$g = \min\{i : \text{Ext}^i_R(R/I, B) \neq \{0\}\}.$$

Proof. Let $x_1, \ldots, x_g$ be a maximal B-sequence in I. For all $i = 1, 2, \ldots, g$, define $I_i = (x_1, \ldots, x_{i-1})$ (with $I_1 = \{0\}$). Now x_i is a (B/I_iB)-regular element, and so

$$\text{Ext}^{i-1}_R(R/I, B) \cong \text{Hom}_R(R/I, B/I_iB) = \{0\},$$

by Lemma 11.179, which applies because $\text{ann}(R/I) = I \supseteq I_i$. On the other hand, since $x_1, \ldots, x_g$ is a maximal B-sequence in I, the ideal I contains no (B/IB)-regular elements. Thus, Lemma 11.178 gives

$$\text{Ext}^g_R(R/I, B) \cong \text{Hom}_R(R/I, B/IB) \neq \{0\}. \quad \bullet$$

Definition. If R is a noetherian ring, B a finitely generated R-module, and I an ideal such that $IB \neq B$, then the ***grade*** of B in I is

$$G(I, B) = \text{length of a maximal } B\text{-sequence in } I.$$

If $(R, \mathfrak{m})$ is a noetherian local ring, then $G(\mathfrak{m}, B)$ is called ***depth*** of B:

$$\text{depth}(B) = G(\mathfrak{m}, B).$$

The number d in Lemma 11.176 is $\text{depth}(R)$.

Proposition 11.181 (Auslander–Buchsbaum). *Let $(R, \mathfrak{m})$ be a noetherian local ring, and let B be a finitely generated R-module with $pd(B) = n < \infty$. Then*

$$pd(B) + \operatorname{depth}(B) = \operatorname{depth}(R).$$

Proof. The proof is by induction on $n = pd(B) \geq 0$. If $n = 0$, then B is a finitely generated projective R-module, and so B is free, by Proposition 11.23. Hence, $B \cong \sum_{j=1}^{m} R_j$, where $R_j \cong R$, and so $\operatorname{Ext}^q_R(k, B) \cong \sum_{j=1}^{m} \operatorname{Ext}^q_R(k, R)$ for all q. It follows that $\operatorname{depth}(B) = \operatorname{depth}(R)$, as desired.

For the inductive step, there is an exact sequence

$$0 \to \Omega \to F \to B \to 0,$$

where F is a finitely generated free R-module. The long exact sequence is

$$\operatorname{Ext}^i_R(k, F) \to \operatorname{Ext}^i_R(k, B) \to \operatorname{Ext}^{i+1}_R(k, \Omega) \to \operatorname{Ext}^{i+1}_R(k, F).$$

By Lemma 11.178, $\operatorname{Ext}^0_R(k, F) = \operatorname{Hom}_R(k, F) = \{0\}$; since F is free, $\operatorname{Ext}^i_R(k, F) = \{0\}$ for all $i > 0$. Therefore, $\operatorname{Ext}^i_R(k, B) \cong \operatorname{Ext}^{i+1}_R(k, \Omega)$ for all $i \geq 0$. It follows that

$$\operatorname{depth}(\Omega) = \operatorname{depth}(B) + 1.$$

Since $n = pd(B) > 0$, we have B not projective, and so $pd(\Omega) = n - 1$. By induction, $pd(\Omega) + \operatorname{depth}(\Omega) = \operatorname{depth}(R)$. Therefore,

$$\begin{aligned}\operatorname{depth}(R) &= pd(\Omega) + \operatorname{depth}(\Omega)\\ &= pd(\Omega) + 1 + \operatorname{depth}(\Omega) - 1\\ &= pd(B) + \operatorname{depth}(B). \quad \bullet\end{aligned}$$

Corollary 11.182. *If $(R, \mathfrak{m})$ is a noetherian local ring of finite global dimension, then*

$$D(R) \leq \operatorname{depth}(R).$$

Proof. By Proposition 11.181, $pd(M) \leq \operatorname{depth}(R)$ for every finitely generated R-module M. But $D(R) = \sup\{pd(M) : M \text{ is finitely generated}\}$, by Theorem 11.134, and so $D(R) \leq \operatorname{depth}(R)$. $\bullet$

To complete the homological characterization of regular local rings, it remains to establish the second inequality in Lemma 11.176: $\mu(\mathfrak{m}) \leq D(R)$. Recall that Theorem 11.157 shows that $D(R) = pd(k)$. If we write $s = \mu(\mathfrak{m})$, we will prove that $s \leq pd(k)$ by comparing a *Koszul complex* for k with a *minimal resolution* of k. Our exposition is merely a more detailed version of the account given in Serre, *Algèbre Locale: Multiplicités*, pages 112–116.

If $f : A \to B$ is a map of R-modules, let $\overline{f} = f \otimes 1_k : A \otimes_R k \to B \otimes_R k$. Recall that $A \otimes_R k \cong A/\mathfrak{m}A$, as can easily be seen by tensoring the short exact sequence $0 \to \mathfrak{m} \to R \to k \to 0$ by A. With this identification, $\overline{f} : a + \mathfrak{m}A \mapsto f(a) + \mathfrak{m}B$.

Lemma 11.183. *Let $(R, \mathfrak{m}, k)$ be a noetherian local ring, let $f: A \to B$ be a map of finitely generated R-modules, and let $\overline{f} = f \otimes 1_k$.*

(i) *$\overline{f}$ is surjective if and only if f is surjective.*

(ii) *If, in addition, both A and B are free R-modules, then $\overline{f}$ injective implies that f is a (split) injection.*

Proof. (i) If $\overline{f}$ is surjective, tensor the exact sequence $A \xrightarrow{f} B \to \operatorname{coker} f \to 0$ by k to obtain the exact sequence

$$A \otimes_R k \xrightarrow{\overline{f}} B \otimes_R k \to (\operatorname{coker} f) \otimes_R k \to 0.$$

Since $\overline{f}$ is surjective, $(\operatorname{coker} f) \otimes_R k = \{0\}$. But $(\operatorname{coker} f) \otimes_R k \cong \operatorname{coker} f/\mathfrak{m} \operatorname{coker} f$, so that $\operatorname{coker} f = \mathfrak{m} \operatorname{coker} f$. Now $\operatorname{coker} f$ is finitely generated, because B is finitely generated, and so Nakayama's lemma gives $\operatorname{coker} f = \{0\}$; that is, f is surjective.

Conversely, since $\otimes_R k$ is right exact, f surjective implies $\overline{f}$ surjective.

(ii) Assume that $\overline{f}$ is injective. Let $x_1, \ldots, x_t$ be a basis of A, and let $b_i = f(x_i)$ for $i = 1, \ldots, t$. Since $\overline{f}$ is injective, the elements $\overline{b}_i = b_i + \mathfrak{m}B$ are linearly independent in $B/\mathfrak{m}B$, and so they extend to a basis: There are $c_1, \ldots, c_s \in B$ with $\overline{b}_1, \ldots, \overline{b}_t, \overline{c}_1, \ldots, \overline{c}_s$ a basis of $B/\mathfrak{m}B$. An application of Nakayama's lemma, as in the proof of Proposition 11.23, shows that $b_1, \ldots, b_t, c_1, \ldots, c_s$ is a basis of B. If we define $h: B \to A$ by $h(b_i) = x_i$ and $h(c_j) = 0$, then we see that $hf = 1_A$, and so f is injective. •

Definition. Let $(R, \mathfrak{m}, k)$ be a noetherian local ring. A map $f: A \to B$ of R-modules is ***minimal*** if $\ker f \subseteq \mathfrak{m}A$.

Thus, the lemma says that if $f: A \to B$ is minimal, where A and B are free R-modules of finite rank, then $\overline{f}: \overline{A} \to \overline{B}$ injective implies f injective.

Definition. Let $(R, \mathfrak{m}, k)$ be a noetherian local ring, and let A be a finitely generated R-module. A free resolution

$$\cdots \to L_2 \xrightarrow{d_2} L_1 \xrightarrow{d_1} L_0 \xrightarrow{d_0} A \to 0$$

is a ***minimal resolution*** if all L_n are finitely generated and $\ker d_n \subseteq \mathfrak{m}L_n$ for all $n \geq 0$; that is, all d_n are minimal.

Proposition 11.184. *Let $(R, \mathfrak{m}, k)$ be a noetherian local ring. Every finitely generated R-module A has a minimal resolution.*

Proof. Since A is finitely generated, it has a minimal generating set, say, $\{a_1, \ldots, a_n\}$. Let L_0 be the free R-module with basis $\{e_1, \ldots, e_n\}$, and define $d_0: L_0 \to A$ by $d_0(e_i) = a_i$ for all i. We saw, in the proof of Proposition 11.23 that $\ker d_0 \subseteq \mathfrak{m}L_0$, and so d_0 is minimal. Since R is noetherian, $\ker d_0$ is finitely generated, and so this construction can be iterated. Thus, induction shows that a minimal resolution of A exists. •

Proposition 11.185. *Let $(R, \mathfrak{m}, k)$ be a noetherian local ring, let A be a finitely generated R-module, and let*

$$\cdots \to L_2 \xrightarrow{d_2} L_1 \xrightarrow{d_1} L_0 \xrightarrow{d_0} A \to 0$$

be a minimal resolution. Then for all $i \geq 0$,

$$\operatorname{Tor}_i^R(A, k) \cong L_i/\mathfrak{m}L_i.$$

Therefore,

$$\operatorname{rank}(\operatorname{Tor}_i^R(A, k)) = \operatorname{rank}(L_i).$$

Proof. Deleting A from the minimal resolution gives a complex $\mathbf{L}_A$; tensoring $\mathbf{L}_A$ by k gives a complex

$$\overline{\mathbf{L}}_A = \cdots \to \overline{L}_2 \xrightarrow{\overline{d}_2} \overline{L}_1 \xrightarrow{\overline{d}_1} \overline{L}_0 \to 0.$$

Now $\operatorname{im} d_{i+1} = \ker d_i \subseteq \mathfrak{m}L_i$ implies $\overline{d}_i = 0$ for every i, so that $H_i(\overline{\mathbf{L}}_A) \cong \overline{L}_i$ for all $i \geq 0$. On the other hand, $\overline{\mathbf{L}}_A = \mathbf{L}_A \otimes_R k$, and so the definition of Tor gives $H_i(\mathbf{L}_A \otimes_R k) = \operatorname{Tor}_i^R(A, k)$. Therefore, $\operatorname{Tor}_i^R(A, k) \cong \overline{L}_i \cong L_i/\mathfrak{m}L_i$. •

Let us make an elementary observation. If $(R, \mathfrak{m}, k)$ is a noetherian local ring and M is an R-module, then we have already seen that $M/\mathfrak{m}M \cong M \otimes_R k$; let us denote $M/\mathfrak{m}M$ by $\overline{M}$. A map $\varphi \colon M \to M'$ induces a map $\overline{\varphi} \colon \overline{M} \to \overline{M}'$ by

$$\overline{\varphi} \colon u + \mathfrak{m}M \mapsto \varphi(u) + \mathfrak{m}M';$$

if φ satisfies the additional condition $\operatorname{im} \varphi \subseteq \mathfrak{m}M'$, then there is a second induced map $\widetilde{\varphi} \colon M/\mathfrak{m}M \to \mathfrak{m}M/\mathfrak{m}^2M$, given by

$$\widetilde{\varphi}(u + \mathfrak{m}M) = \varphi(u) + \mathfrak{m}^2M'.$$

Lemma 11.186. *Let $(R, \mathfrak{m}, k)$ be a local ring, let A be a finitely generated R-module, and let*

$$\cdots \to L_2 \xrightarrow{d_2} L_1 \xrightarrow{d_1} L_0 \xrightarrow{d_0} A \to 0$$

be a minimal resolution of A. If $\cdots \to M_2 \xrightarrow{D_2} M_1 \xrightarrow{D_1} M_0 \xrightarrow{\varepsilon} A \to 0$ is a complex satisfying

(i) *each M_p is a finitely generated free R-module*;

(ii) *$\overline{\varepsilon} \colon \overline{M}_0 \to \overline{A}$ is injective*;

(iii) *For all $p > 0$, we have $D_p(M_p) \subseteq \mathfrak{m}M_{p-1}$, and $\widetilde{D}_p \colon \overline{M}_p \to \mathfrak{m}M_{p-1}/\mathfrak{m}^2M_{p-1}$, given by $u_p + \mathfrak{m}M_p \mapsto D_p(u_p) + \mathfrak{m}^2M_{p-1}$, is an injection*;

then, for all $p \geq 0$, $\operatorname{rank}(M_p) \leq \operatorname{rank}(L_p) = \operatorname{rank}(\operatorname{Tor}_p^R(A, k))$.

Proof. We will show that each M_p is isomorphic to a direct summand of L_p. By Theorem 10.46, the comparison theorem, there are maps f_p making the following diagram commute:

$$\begin{array}{ccccccccc}
\cdots & \longrightarrow & M_1 & \xrightarrow{D_1} & M_0 & \xrightarrow{\varepsilon} & A & \longrightarrow & 0 \\
 & & \downarrow{\scriptstyle f_1} & & \downarrow{\scriptstyle f_0} & & \downarrow{\scriptstyle 1_A} & & \\
\cdots & \longrightarrow & L_1 & \xrightarrow[d_1]{} & L_0 & \xrightarrow[d_0]{} & A & \longrightarrow & 0.
\end{array}$$

It suffices to find surjections $g_p \colon L_p \to M_p$: since M_p is free, hence projective, M_p would then be isomorphic to a direct summand of L_p. We claim that such maps g_p exist if $\overline{f}_p \colon \overline{M}_p \to \overline{L}_p$ is injective. Now $\overline{M}_p$ and $\overline{L}_p$ are vector spaces over k, so that the subspace $\overline{f}_p(\overline{M}_p) \cong \overline{M}_p$ is a direct summand of $\overline{L}_p$; thus, there is a (necessarily) surjective map $\gamma \colon \overline{L}_p \to \overline{M}_p$ with $\gamma \overline{f}_p = 1_{\overline{M}_p}$. Let $\pi \colon M_p \to \overline{M}_p$ and $\nu \colon L_p \to \overline{L}_p$ be the natural maps (regard $\overline{M}_p = M_p/\mathfrak{m}M_p$ and $\overline{L}_p = L_p/\mathfrak{m}L_p$), and consider the diagram

$$\begin{array}{ccccc}
 & & L_p & & \\
 & {\scriptstyle g_p}\swarrow & \downarrow{\scriptstyle \gamma\pi} & & \\
M_p & \xrightarrow[\nu]{} & \overline{M}_p & \longrightarrow & 0.
\end{array}$$

Since L_p is free, there exists g_p with $\nu g_p = \gamma\pi$; that is, $\overline{g}_p = \gamma\pi$. Hence, $\overline{g}_p$ is surjective, and so g_p is surjective, by Lemma 11.183.

It remains to show, by induction on $p \geq 0$, that the conditions listed in the statement imply each $\overline{f}_p$ is injective. For the base step, $d_0 f_0 = \varepsilon$ implies $\overline{d_0}\,\overline{f}_0 = \overline{\varepsilon}$. By hypothesis, both $\overline{\varepsilon}$ and $\overline{d_0}$ are injections. However, Lemma 11.183(i) shows that both are isomorphisms, because both ε and d_0 are surjections. It follows that $\overline{f}_0$ is an injection (in fact, it is even an isomorphism).

For the inductive step, consider the following commutative diagram.

$$\begin{array}{ccc}
M_p/\mathfrak{m}M_p & \xrightarrow{\overline{f}_p} & L_p/\mathfrak{m}L_p \\
{\scriptstyle \widetilde{D}_p}\downarrow & & \downarrow{\scriptstyle \widetilde{d}_p} \\
\mathfrak{m}M_{p-1}/\mathfrak{m}^2M_{p-1} & \xrightarrow[\overline{f}_{p-1}]{} & \mathfrak{m}L_{p-1}/\mathfrak{m}^2L_{p-1}.
\end{array}$$

Since $\mathbf{L}_\bullet$ is a complex, we have $\operatorname{im} d_p \subseteq \ker d_{p-1}$; since $\mathbf{L}_\bullet$ is a minimal resolution, we have $\ker d_{p-1} \subseteq \mathfrak{m}L_{p-1}$; hence, the map $\widetilde{d}_p$ is defined. By induction, $\overline{f}_{p-1}$ is injective; hence, $\overline{f}_{p-1}\widetilde{D}_p$ is injective, because $\widetilde{D}_p$ is injective, by hypothesis. Therefore, $\widetilde{d}_p\overline{f}_p$ is injective, and this implies that $\overline{f}_p$ is injective. •

The following complex $\mathbf{M}_\bullet$ will be seen to satisfy the conditions in Lemma 11.186.

Definition. Let $x_1, \ldots, x_s$ be a sequence of elements in a commutative ring R. The ***Koszul complex*** $\mathbf{M}(x_1, \ldots, x_s)_\bullet$ is defined as follows.

$$\mathbf{M}(x_1, \ldots, x_s)_p = \bigwedge^p(F),$$

where F is the free R-module with basis $\{e_1, \ldots, e_s\}$. The differentiations $D_p\colon \bigwedge^p(F) \to \bigwedge^{p-1}(F)$ are defined by

$$D_1\bigl(\sum_{i=1}^{s} c_i e_i\bigr) = \sum_{i=1}^{s} c_i x_i,$$

where $c_i \in R$ for all i (so that $D_1(e_i) = x_i$), and, for $p > 1$,

$$D_p(e_{i_1} \wedge \cdots \wedge e_{i_p}) = \sum_{r=0}^{p} (-1)^{r-1} x_r e_{i_1} \wedge \cdots \wedge \widehat{e_{i_r}} \wedge \cdots \wedge e_{i_p}.$$

If A is an R-module, the ***Koszul complex*** $\mathbf{M}(x_1, \ldots, x_s, A)_\bullet$ is defined by

$$\mathbf{M}(x_1, \ldots, x_s, A)_\bullet = A \otimes_R \mathbf{M}(x_1, \ldots, x_s)_\bullet.$$

We leave the straightforward calculation that $D_{p-1}D_p = 0$ to the reader; it is similar to that in the proof of Lemma 10.114. Thus, the Koszul complex really is a complex.

Note that $\bigwedge^0(F) = R$ and that $\operatorname{im} d_1 = I$, where $I = (x_1, \ldots, x_s)$. In general, the Koszul complex is not acyclic; that is, it is not an exact sequence. However, if $x_1, \ldots, x_s$ is an R-sequence, then augmenting it with the natural map $\varepsilon\colon \bigwedge^0(F) \to R/I$ gives a free resolution of R/I (see Bruns–Herzog, *Cohen–Macaulay Rings*, page 49).

Observe that the pth term of $\mathbf{M}(x_1, \ldots, x_s, k)_\bullet$ is, by definition, $k \otimes_R \bigwedge^p(F)$. Since F is free of rank s, we know, from Theorem 9.140, the binomial theorem, that $\bigwedge^p(F)$ is free of rank $\binom{s}{p}$, and so $k \otimes_R \bigwedge^p(F)$ is a vector space over k of dimension $\binom{s}{p}$. Thus, if we denote $\bigwedge^p(F)$ by M_p, as in Lemma 11.186, then $k \otimes_R \bigwedge^p(F)$ is $\overline{M}_p$.

If $x_1, \ldots, x_s$ is a minimal generating set for $\mathfrak{m}$, then Proposition 11.165 says that $x_1^*, \ldots, x_s^*$ is a basis for $\mathfrak{m}/\mathfrak{m}^2$, where $x_i^* = x_i + \mathfrak{m}^2$. If M is an R-module, then there is an isomorphism $\mathfrak{m}M/\mathfrak{m}^2M \to (\mathfrak{m}/\mathfrak{m}^2) \otimes_R M$, given by

$$\sum_i x_i v_i + \mathfrak{m}^2 M \mapsto \sum_i x_i^* \otimes v_i,$$

where $v_i \in M$. If $\varphi\colon M \to M'$ has the property that $\operatorname{im}\varphi \subseteq \mathfrak{m}M'$, then $\varphi(u) = \sum_i x_i v_i'$, where $v_i' \in M'$. Composing $\widetilde{\varphi}\colon M/\mathfrak{m}M \to \mathfrak{m}M'/\mathfrak{m}^2M'$ with the isomorphism above allows us to regard $\widetilde{\varphi}\colon M/\mathfrak{m}M \to (\mathfrak{m}/\mathfrak{m}^2) \otimes_R M'$:

$$\widetilde{\varphi}\colon u + \mathfrak{m}M \mapsto \varphi(u) + \mathfrak{m}^2M' = \sum_i x_i v_i' + \mathfrak{m}^2 M' \mapsto \sum_i x_i^* \otimes v_i'.$$

Lemma 11.187. *Let $(R, \mathfrak{m}, k)$ be a noetherian local ring, and let $x_1, \ldots, x_s$ be a minimal generating set for $\mathfrak{m}$. Then the Koszul complex $\mathbf{M}(x_1, \ldots, x_s)_\bullet$ satisfies the conditions in* Lemma 11.186.

Proof. First, each term $M_p = \bigwedge^p(F)$ of the Koszul complex is a finitely generated free R-module. Second, define $\varepsilon\colon R \to k$ to be the natural map. Since $\ker\varepsilon = \mathfrak{m}$, the map $\overline{\varepsilon}$ is an injection, by Lemma 11.183. For the third condition, recall the formula for $D_p\colon \bigwedge^p(F) \to \bigwedge^{p-1}(F)$ (where F is the free R-module with basis $e_1, \ldots, e_s$):

$$D_p(e_{i_1} \wedge \cdots \wedge e_{i_p}) = \sum_{r=1}^{p} (-1)^{r-1} x_r e_{i_1} \wedge \cdots \wedge \widehat{e_{i_r}} \wedge \cdots \wedge e_{i_p}.$$

The presence of the factor x_r forces each term, and hence the sum, into $\mathfrak{m}M_{p-1}$.

Finally, if $M_p = \bigwedge^p(F)$ and $M_{p-1} = \bigwedge^{p-1}(F)$, let us show that $\widetilde{D}_p\colon M_p/\mathfrak{m}M_p \to \mathfrak{m}M_{p-1}/\mathfrak{m}^2 M_{p-1}$, given by

$$\widetilde{D}_p(u + \mathfrak{m}M_p) = D_p(u) + \mathfrak{m}^2 M_{p-1},$$

is an injection. Recall that if $e_1, \ldots, e_s$ is a basis of the free module F, then a basis for $M_p = \bigwedge^p(F)$ is the set of all e_I, where $I = i_1 < \cdots < i_p$ is an increasing $p \le s$ list and $e_I = e_{i_1} \wedge \cdots \wedge e_{i_p}$. If $u = \sum_I \alpha_I e_I$, we may assume that α_I is defined for every increasing list I (some α_I may be 0). Let us now define α_I for every, not necessarily increasing, p-tuple $i_1, \ldots, i_p$ of indices, possibly with a repeated index: set $\alpha_I = 0$ if some index is repeated, and set $\alpha_I = -\alpha_{I'}$ if I' is obtained from I by transposing two indices. With this notation, we may now rewrite the formula for D_p:

$$\begin{aligned} D_p(u) &= D_p\Big(\sum_I \alpha_I e_I\Big) \\ &= \sum_{j=1}^{s} x_j \sum_L \alpha_{jL} e_L, \end{aligned}$$

where $L = \ell_1 < \cdots < \ell_{p-1}$ is an increasing $p - 1 \le s$ list, and $jL = j, \ell_1, \ldots, \ell_{p-1}$. Note that α_{jL} is either 0 or $\pm\alpha_I$, where I is the rearrangement of jL into an increasing list. Suppose now that $u = \sum_I \alpha_I e_I \notin \mathfrak{m}M_p$; that is, $u + \mathfrak{m}M_p \ne 0$ in $M_p/\mathfrak{m}M_p$. Since the e_Is are a basis of M_p, we must have some $\alpha_I \notin \mathfrak{m}$; that is, α_I is a unit in R. If $I = i_1 < \cdots < i_p$, define $j = i_1$ and $L = i_2 < \cdots < i_p$. The coefficient $\alpha_{jL} = \alpha_I$ of e_L does not lie in $\mathfrak{m}$, and so $\sum_L \alpha_{jL} e_L \ne 0$ in M_{p-1} (because the e_Ls form a basis of M_{p-1}). Under the isomorphism $\mathfrak{m}M_{p-1}/\mathfrak{m}^2 M_{p-1} \to (\mathfrak{m}/\mathfrak{m}^2) \otimes_R M_{p-1}$,

$$\widetilde{D}_p(u) = \sum_{j=1}^{s} x_j^* \otimes \sum_L \alpha_{jL} e_L,$$

where $x_j^* = x_j + \mathfrak{m}$. Since $x_i^*, \ldots, x_s^*$ is a basis of $\mathfrak{m}/\mathfrak{m}^2$, an element $\sum_j x_j^* \otimes v_j = 0$ if and only if each $v_j = 0$. Therefore, if $u \notin \mathfrak{m}M_p$, then $\widetilde{D}_p(u + \mathfrak{m}M_p) \ne 0$, and so $\widetilde{D}_p$ is an injection. •

Proposition 11.188. *If $(R, \mathfrak{m}, k)$ is a noetherian local ring of finite global dimension $D(R)$, then*

$$\mu(\mathfrak{m}) \leq D(R).$$

Proof. Let $s = \mu(\mathfrak{m})$, and let $\{x_1, \ldots, x_s\}$ be a minimal generating set of $\mathfrak{m}$. Then

$$\operatorname{rank}(\mathbf{M}(x_1, \ldots, x_s)_p) \leq \operatorname{rank}(\operatorname{Tor}_p^R(k, k)),$$

by Lemma 11.186. Now $\mathbf{M}(x_1, \ldots, x_s)_p = \bigwedge^p(F)$, where F is the free R-module with basis $e_1, \ldots, e_s$, so that $\operatorname{rank}(\mathbf{M}(x_1, \ldots, x_s)_p) = \operatorname{rank}(\bigwedge^p(F)) = \binom{s}{p}$ and

$$\binom{s}{p} \leq \operatorname{rank}(\operatorname{Tor}_p^R(k, k)).$$

Therefore, $1 \leq \operatorname{rank}(\operatorname{Tor}_s^R(k, k))$, so that $\operatorname{Tor}_s^R(k, k) \neq \{0\}$. But Lemma 11.155 gives

$$pd(k) = \max\big\{p : \operatorname{Tor}_p^R(k, k) \neq \{0\}\big\},$$

so that $s \leq pd(k) = D(R)$, by Corollary 11.156. •

Theorem 11.189 (Serre). *A noetherian local ring R is regular if and only if $D(R)$ is finite.*

Proof. In Lemma 11.176, the theorem was reduced to checking two inequalities. These inequalities are proved in Corollary 11.182 and Proposition 11.188. •

Corollary 11.190. *If R is a regular local ring, and it $\mathfrak{p}$ is a prime ideal in R, then $R_\mathfrak{p}$ is also a regular local ring.*

Proof. Since $\mathfrak{p}$ is a prime ideal, the localization $R_\mathfrak{p}$ is a local ring; it is noetherian because R is noetherian. In Proposition 11.154, we saw that $D(R) \geq D(R_\mathfrak{p})$. Therefore, $R_\mathfrak{p}$ is a regular local ring, by Serre's theorem. •

We are now going to prove that every regular local ring is a UFD, and we begin with several elementary lemmas.

Lemma 11.191. *If R is a noetherian domain, then R is a* UFD *if and only if every prime ideal of height* 1 *is principal.*

Proof. Let R be a UFD, and let $\mathfrak{p}$ be a prime ideal of height 1. If $a \in \mathfrak{p}$ is nonzero, then $a = p_1^{e_1} \cdots p_n^{e_n}$, where the p_i are irreducible and $e_i \geq 1$. Since $\mathfrak{p}$ is prime, one of the factors, say, $p_j \in \mathfrak{p}$. Of course, $Rp_j \subseteq \mathfrak{p}$. But Rp_j is a prime ideal, by Proposition 6.17, so that $Rp_j = \mathfrak{p}$, because $\operatorname{ht}(\mathfrak{p}) = 1$.

Conversely, since R is noetherian, Lemma 6.18 shows that every nonzero nonunit in R is a product of irreducibles, and so Proposition 6.17 says that it suffices to prove, for every irreducible $\pi \in R$, that $R\pi$ is a prime ideal. Choose a prime ideal $\mathfrak{p}$ that is minimal over $R\pi$. By the principal ideal theorem, Theorem 11.161, we have $\operatorname{ht}(\mathfrak{p}) = 1$, and so the hypothesis gives $\mathfrak{p} = Ra$ for some $a \in R$. Therefore, $\pi = ua$ for some $u \in R$. Since π is irreducible, we must have u a unit, and so $R\pi = Ra = \mathfrak{p}$, as desired. •

Lemma 11.192. *Let R be a noetherian domain, let $x \in R$ be a nonzero element with Rx a prime ideal, and denote $S^{-1}R$ by R_x, where $S = \{x^n : n \geq 0\}$. Then R is a* UFD *if and only if R_x is a* UFD.

Proof. We leave necessity as an exercise for the reader. For sufficiency, assume that R_x is a UFD. Let $\mathfrak{p}$ be a prime ideal in R of height 1. If $x \in \mathfrak{p}$, then $Rx \subseteq \mathfrak{p}$ and, since $\mathrm{ht}(\mathfrak{p}) = 1$, we have $Rx = \mathfrak{p}$ (for Rx is prime), and so $\mathfrak{p}$ is principal in this case. We may now assume that $x \notin \mathfrak{p}$; that is, $S \cap \mathfrak{p} = \varnothing$. It follows that $\mathfrak{p}R_x$ is a prime ideal in R_x of height 1, and so it is principal, by hypothesis. There is some $a \in \mathfrak{p}$ and $n \geq 0$ with $\mathfrak{p}R_x = R_x(a/x^n) = R_x a$, for x is a unit in R_x. We may assume that $a \notin Rx$. If $a = a_1 x$ and $a_1 \notin Rx$, then replace a by a_1, for $R_x a = R_x a_1$. If $a_1 = a_2 x$ and $a_2 \notin Rx$, then replace a_1 by a_2, for $R_x a_1 = R_x a_2$. If this process does not stop, there are equations $a_m = a_{m+1}x$ for all $m \geq 1$, which give rise to an ascending sequence $Ra_1 \subseteq Ra_2 \subseteq \cdots$. Since R is noetherian, $Ra_m = Ra_{m+1}$ for some m. Hence, $a_{m+1} = ra_m$ for some $r \in R$, and $a_m = a_{m+1}x = ra_m x$. Since R is a domain, $1 = rx$; thus, x is a unit, contradicting Rx being a prime (hence, proper) ideal. Clearly, $Ra \subseteq \mathfrak{p}$; we claim that $Ra = \mathfrak{p}$. If $b \in \mathfrak{p}$, then $b = (r/x^m)a$ in R_x, where $r \in R$ and $m \geq 0$. Hence, $x^m b = ra$ in R. Choose m minimal. If $m > 0$, then $ra = x^m b \in Rx$; since Rx is prime, either $r \in Rx$ or $a \in Rx$. But $a \notin Rx$ since $S \cap \mathfrak{p} = \varnothing$, so that $r = xr'$. As R is a domain, this gives $r'a = x^{m-1}b$, contradicting the minimality of m. We conclude that $m = 0$, and so $\mathfrak{p} = Ra$ is principal. Lemma 11.191 now shows that R is a UFD. •

The following elementary lemma is true when the localizing ideal is prime; however, we will use it only in the case the ideal is maximal.

Lemma 11.193. *Let R be a domain, and let I be a nonzero projective ideal in R. If $\mathfrak{m}$ is a maximal ideal in R, then*

$$I_{\mathfrak{m}} \cong R_{\mathfrak{m}}.$$

Proof. Since I is a projective R-module, $I_{\mathfrak{m}}$ is a projective $R_{\mathfrak{m}}$-module. As $R_{\mathfrak{m}}$ is a local ring, however, $I_{\mathfrak{m}}$ is a free $R_{\mathfrak{m}}$-module. But $I_{\mathfrak{m}}$ is an ideal in a domain $R_{\mathfrak{m}}$, and so it must be principal; that is, $I_{\mathfrak{m}} \cong R_{\mathfrak{m}}$. •

Theorem 11.194 (Auslander–Buchsbaum). *Every regular local ring R is a* UFD.

Proof. (***Kaplansky***) The proof is by induction on the Krull dimension $\dim(R)$, the cases $n = 0$ (R is a field) and $n = 1$ (R is a DVR) being obvious (see Exercise 11.78 on page 1012). For the inductive step, choose $x \in \mathfrak{m} - \mathfrak{m}^2$. By Lemma 11.171, R/Rx is a regular local ring with $\dim(R/Rx) < \dim(R)$; by Proposition 11.172, R/Rx is a domain, and so Rx is a prime ideal. It suffices, by Lemma 11.192, to prove that R_x is a UFD (where $R_x = S^{-1}R$ for $S = \{x^n : n \geq 0\}$). Let $\mathfrak{P}$ be a prime ideal of height 1 in R_x; we must show that $\mathfrak{P}$ is principal. Define $\mathfrak{p} = \mathfrak{P} \cap R$ (since R is a domain, $R_x \subseteq \mathrm{Frac}(R)$, so that the intersection makes sense). Since R is a regular local ring, $D(R) < \infty$, and so the R-module $\mathfrak{p}$ has a free resolution of finite length:

$$0 \to F_n \to F_{n-1} \to \cdots \to F_0 \to \mathfrak{p} \to 0.$$

Tensoring by R_x, which is a flat R-module (Theorem 11.28), gives a free R_x-resolution of $\mathfrak{P}$ (for $\mathfrak{P} = R_x\mathfrak{p}$):

$$0 \to F'_n \to F'_{n-1} \to \cdots \to F'_0 \to \mathfrak{P} \to 0, \tag{9}$$

where $F'_i = R_x \otimes_R F_i$.

We claim that $\mathfrak{P}$ is projective. By Proposition 11.154, it suffices to show that every localization $\mathfrak{P}_{\mathfrak{M}}$ is projective, where $\mathfrak{M}$ is a maximal ideal in R_x. Now $(R_x)_{\mathfrak{M}}$ is a localization of R, and so it is a regular local ring, by Corollary 11.190; its dimension is smaller than $D(R)$, and so it is a UFD, by induction. Now $\mathfrak{P}_{\mathfrak{M}}$, being a height 1 prime ideal in the UFD $(R_x)_{\mathfrak{M}}$, is principal. But principal ideals in a domain are free, hence projective, and so $\mathfrak{P}_{\mathfrak{M}}$ is projective. Therefore, $\mathfrak{P}$ is projective.

The exact sequence (9) "factors" into split short exact sequences. Since $\mathfrak{P}$ is projective, we have $F'_0 \cong \mathfrak{P} \oplus \Omega_0$, where $\Omega_0 = \ker(F'_0 \to \mathfrak{P})$. Thus, Ω_0 is projective, being a summand of a free module, and so $F'_1 \cong \Omega_1 \oplus \Omega_0$, where $\Omega_1 = \ker(F'_1 \to F'_0)$. More generally, $F'_i \cong \Omega_i \oplus \Omega_{i-1}$ for all $i \geq 1$. Hence,

$$F'_0 \oplus F'_1 \oplus \cdots \oplus F'_n \cong (\mathfrak{P} \oplus \Omega_0) \oplus (\Omega_1 \oplus \Omega_0) \oplus \cdots .$$

Since projective modules over a local ring are free, we see that there are finitely generated free R_x-modules Q and Q' with

$$Q \cong \mathfrak{P} \oplus Q'.$$

Recall that $\operatorname{rank}(Q) = \dim_K(K \otimes_{R_x} Q)$, where $K = \operatorname{Frac}(R_x)$; now $\operatorname{rank}(\mathfrak{P}) = 1$ and $\operatorname{rank}(Q') = r$, say, so that $\operatorname{rank}(Q) = r + 1$.

We must still show that $\mathfrak{P}$ is principal. Now

$$\bigwedge^{r+1}(Q) \cong \bigwedge^{r+1}(\mathfrak{P} \oplus Q').$$

Since Q is free of rank $r+1$, Theorem 9.140, the binomial theorem, gives $\bigwedge^{r+1}(Q) \cong R_x$. On the other hand, Theorem 9.143 gives

$$\bigwedge^{r+1}(\mathfrak{P} \oplus Q') \cong \sum_{i=0}^{r+1}\Big(\bigwedge^{i}(\mathfrak{P}) \otimes_{R_x} \bigwedge^{r+1-i}(Q')\Big). \tag{10}$$

We claim that $\bigwedge^i(\mathfrak{P}) = \{0\}$ for all $i > 1$. By Lemma 11.193, we have $\mathfrak{P}_{\mathfrak{M}} \cong (R_x)_{\mathfrak{M}}$ for every maximal ideal $\mathfrak{M}$ in R_x. Now Exercise 11.24 on page 921 gives

$$\Big(\bigwedge^{i}(\mathfrak{P})\Big)_{\mathfrak{M}} \cong \bigwedge^{i}(\mathfrak{P}_{\mathfrak{M}}) \cong \bigwedge^{i}((R_x)_{\mathfrak{M}})$$

for all maximal ideals $\mathfrak{M}$ and all i. But $\bigwedge^i((R_x)_{\mathfrak{M}}) = \{0\}$ for all $i > 1$ (by the binomial theorem or by the simpler Corollary 9.138), so that Proposition 11.31 gives $\bigwedge^i(\mathfrak{P}) = \{0\}$ for all $i > 1$.

We have just seen that most of the terms in (10) are $\{0\}$; what survives is:

$$\bigwedge^{r+1}(\mathfrak{P} \oplus Q') \cong \Big(\bigwedge^{0}(\mathfrak{P}) \otimes_{R_x} \bigwedge^{r+1}(Q')\Big) \oplus \Big(\bigwedge^{1}(\mathfrak{P}) \otimes_{R_x} \bigwedge^{r}(Q')\Big).$$

But $\bigwedge^{r+1}(Q') = \{0\}$ and $\bigwedge^{r}(Q') \cong R_x$, because Q' is free of rank r. Therefore, $\bigwedge^{r+1}(\mathfrak{P} \oplus Q') \cong \mathfrak{P}$. Since $\mathfrak{P} \cong \bigwedge^{r+1}(\mathfrak{P} \oplus Q') \cong \bigwedge^{r+1}(Q) \cong R_x$, we have $\mathfrak{P} \cong R_x$ is principal. Thus, R_x, and hence R, is a UFD. •

Having studied localization, we turn, briefly, to globalization, merely describing its setting. To a given commutative noetherian ring R, we have associated a family of local rings $R_\mathfrak{p}$, one for each prime ideal $\mathfrak{p}$. Local rings are simpler than general rings; for example, if R is a Dedekind ring, its localizations are all principal ideal domains. Globalization asks how we can make use of *all* the localizations to gather information. Consider the disjoint union

$$E(R) = \bigcup_{\mathfrak{p} \in \operatorname{Spec}(R)} R_\mathfrak{p}.$$

We call $R_\mathfrak{p}$ the ***stalk*** of R over $\mathfrak{p}$, and we define the ***projection*** $\pi : E(R) \to \operatorname{Spec}(R)$ by $\pi(e) = \mathfrak{p}$ if $e \in R_\mathfrak{p}$; that is, π sends each point in the stalk over $\mathfrak{p}$ into $\mathfrak{p}$.[8] Each element $a \in R$ defines a function $s_a \colon \operatorname{Spec}(R) \to E(R)$ by

$$s_a : \mathfrak{p} \mapsto a/1 \in R_\mathfrak{p}.$$

Note that $\pi s_a = 1_{\operatorname{Spec}(R)}$. We claim that distinct elements $a, b \in R$ give different functions; that is, if $s_a = s_b$, then $a = b$. Let $I = R(a - b)$. If $(a - b)/1 = 0$ in $R_\mathfrak{p}$ for every prime ideal $\mathfrak{p}$, then $I_\mathfrak{p} = \{0\}$ for every $\mathfrak{p}$. Proposition 11.31 applies to show that $a = b$ (in fact, this proposition only needs $I_\mathfrak{m} = \{0\}$ for all *maximal* ideals $\mathfrak{m}$). Thus, we can regard the elements of any commutative ring R as $E(R)$-valued functions on $\operatorname{Spec}(R)$.

Consider the question, given $f \in R_\mathfrak{p}$ and $g \in R_\mathfrak{q}$, whether there exists $a \in R$ with $f = a/1 \in R_\mathfrak{p}$ and $g = a/1 \in R_\mathfrak{q}$? That is, is there $a \in R$ with $s_a(\mathfrak{p}) = f$ and $s_a(\mathfrak{q}) = g$? A "good" answer might be if $\mathfrak{p}$ and $\mathfrak{q}$ are "close" to each other, then such an element $a \in R$ exists. This suggests that a topology on $X = \operatorname{Spec}(R)$ may be of interest, and the Zariski topology is a good candidate (a subset $F \subseteq \operatorname{Spec}(R)$ is closed if $\mathfrak{q} \in F$ implies $\mathfrak{p} \in F$ whenever $\mathfrak{p}$ is a prime ideal with $\mathfrak{q} \subseteq \mathfrak{p}$). Of course, once we decide on a topology for $\operatorname{Spec}(R)$, we expect that $E(R)$ should also be topologized, and that interesting functions should be continuous.

The Zariski topology is much different from the topology on euclidean space. Not only is it not a metric space (that is, no distance between two points is defined), one-point subsets need not be closed sets; for example, $\{\mathfrak{p}\}$ is closed if and only if $\mathfrak{p}$ is a maximal ideal. In spite of this, ***continuity*** of a function $f \colon X \to Y$ can still be defined: f is continuous if $f^{-1}(V)$ is an open subset of X for every open subset $V \subseteq Y$. Equivalently, f is continuous if and only if, for every closed subset C in Y, the subset $f^{-1}(C)$ is a

[8] There is a possible source of confusion, for $\mathfrak{p}$ can be viewed in two ways: as a prime ideal—a subset of R; as a point in $\operatorname{Spec}(R)$. To distinguish these viewpoints, we often write $\mathfrak{p}_x$ to denote $\mathfrak{p}$ viewed as a point of $X = \operatorname{Spec}(R)$. Thus, the projection π is defined by $\pi(e) = \mathfrak{p}_x$ for all $e \in R_\mathfrak{p}$.

closed subset of X. Similarly, we can define the notions of *compactness* (if $\{F_i : i \in I$ is a family of closed subsets, then there are finitely many of them, say, $F_{i_1}, \ldots, F_{i_n}$ such that $\bigcap_{i \in I} F_i = \bigcap_{j=1}^{n} F_{i_j}$) and *connectedness* (not the union of two nonempty disjoint closed subsets) for arbitrary topological spaces.

Let us mention an elementary ***gluing*** result. Suppose that X and Y are topological spaces, that $\{U_i : i \in I\}$ is a family of open subsets of X, and that $\{f_i : U_i \to Y\}$ is a family of continuous functions. If the functions agree on overlaps; that is, if $f_i|(U_i \cap U_j) = f_j|(U_i \cap U_j)$ for all $i, j \in I$, then there is a unique continuous function $f : \bigcup_i U_i \to Y$ with $f|U_i = f_i$ for all i. Does this remind you of a direct limit?

Here is the formal definition of a sheaf from this viewpoint; there is also a second, equivalent, version using *presheaves* (introduced in Chapter 7) that we will describe afterward.

Definition. If E and X be topological spaces, and let $\pi : E \to X$ be a continuous surjection. Then $\mathcal{F} = (E, X, \pi)$ is a ***sheaf of abelian groups*** if

(i) π is a ***local homeomorphism***: for each $e \in E$, there is an open set U containing e such that $\pi(U)$ is an open subset of X and $\pi|U$ is a homeomorphism[9] from U to $\pi(U)$.

(ii) The subsets $\mathcal{F}_x = \pi^{-1}(x)$ of E, for $x \in X$, are called the ***stalks*** of $\mathcal{F}$, and each $\mathcal{F}_x$ is an abelian group.

(iii) If $E + E$ is the subset of $E \times E$ consisting of all (a, b) with $\pi(a) = \pi(b)$, then the maps $E + E \to E$, given by $(a, b) \mapsto a + b$ and $(a, b) \mapsto a - b$, are continuous.

Ignoring the algebraic structure on the stalks, the reader may recognize the basic ingredients present in covering spaces, for example.

Recalling the functions $s_a : \operatorname{Spec}(R) \to E(R)$ mentioned above, we see that the following notion is of interest.

Definition. If $\mathcal{F} = (E, X, \pi)$ is a sheaf of abelian groups, and if U is an open subset of X, then a ***section over*** U is a continuous function $s : U \to E$ with $\pi s = 1_U$. We write

$$\Gamma(U, \mathcal{F}) = \{\text{all sections } s : U \to E\}.$$

A ***global section*** is a section in $\Gamma(X, \mathcal{F})$.

It is easy to check that $\Gamma(U, \mathcal{F})$ is an abelian group. If $V \subseteq U$ are open subsets of X, then there is a ***restriction map***

$$\rho_V^U : \Gamma(U, \mathcal{F}) \to \Gamma(V, \mathcal{F}),$$

given by $s \mapsto s|V$. Moreover, the functions ρ_V^U are homomorphisms.

[9] A ***homeomorphism*** is a bijection $f : X \to Y$, where X and Y are topological spaces, such that both f and f^{-1} are continuous.

For fixed $x \in X$, let

$$I(x) = \{\text{open sets } U \subseteq X \text{ containing } x\}.$$

It is easy to see that $I(x)$ is a partially ordered set under reverse inclusion:

$$U \preceq V \text{ means } U \supseteq V.$$

In fact, $I(x)$ is a directed index set, for given $U, V \in I(x)$, then $U \cap V \in I(x)$, $U \cap V \subseteq U$, and $U \cap V \subseteq V$; that is, $U \preceq U \cap V$ and $V \preceq U \cap V$. We can recapture the stalks from the sections. Let $\mathcal{F} = (E, X, \pi)$ be a sheaf of abelian groups. For each $x \in X$,

$$\mathcal{F}_x \cong \varinjlim_{U \in I(x)} \Gamma(U, \mathcal{F}).$$

If X is a topological space, then the family of its open sets, with inclusion maps of subsets as morphisms, is a category; denote it by $\mathbf{Open}(X)$. We may now define a *presheaf* of abelian groups. In fact, there are presheaves with values in any category, say, modules or commutative rings, not just **Ab**.

Definition. If X is a topological space and $\mathcal{C}$ is a category, then a ***presheaf*** over X with values in $\mathcal{C}$ is a contravariant functor $\mathcal{F} \colon \mathbf{Open}(X) \to \mathcal{C}$.

A sheaf can be reconstructed from its presheaf of sections if we further assume a version of the gluing result mentioned above.

Given a commutative ring R, sheaves of R-modules are defined whose stalks are $R_\mathfrak{p}$-modules for $\mathfrak{p} \in \operatorname{Spec}(R)$. These form an abelian category with enough injectives; that is, every sheaf can be imbedded as a subsheaf of an injective sheaf. Moreover, global sections $\Gamma(X, \ )$ is a left exact functor, and cohomology of sheaves is defined as derived functors of $\Gamma(X, \ )$. These cohomology groups provide the most important method of globalizing. We recommend the article by J.-P. Serre, *Faisceaux Algèbriques Cohérents*, Annals of Math. (61) 1955, pages 197–278, for a lucid discussion.

EXERCISES

11.75 Let $R = k[x, y, z]$, where k is a field.

(i) Prove that $x, y(1-x), z(1-x)$ is an R-sequence.

(ii) Prove that $y(1-x), z(1-x), x$ is not an R-sequence.

11.76 Let R be a commutative ring. Prove that a prime ideal $\mathfrak{p}$ in R is minimal over an ideal I if and only if $\operatorname{ht}(\mathfrak{p}/I) = 0$ in R/I.

11.77 If $(R, \mathfrak{m}, k)$ is a noetherian local ring, and if B is a finitely generated R-module, prove that

$$\operatorname{depth}(B) = \min\bigl\{i : \operatorname{Ext}^i_R(k, B) \neq \{0\}\bigr\}.$$

11.78 Let R be a regular local ring.

(i) Prove that R is a field if and only if $\dim(R) = 0$.

(ii) Prove that R is a DVR if and only if $\dim(R) = 1$.

11.79 (i) Let $(R, \mathfrak{m})$ be a noetherian local ring, and let $x \in R$ be a regular element; i.e., x is not a zero divisor. If $x_1 + (x), \ldots, x_s + (x)$ is an $(R/(x))$-sequence, prove that $x, x_1, \ldots, x_s$ is an R-sequence.

(ii) Let R be a commutative ring. If $x_1, \ldots, x_d$ is an R-sequence, prove that the cosets $x_2 + (x_1), \ldots, x_d + (x_1)$ form an $(R/(x_1))$-sequence.

11.80 Let R be a noetherian (commutative) ring with Jacobson radical $J = J(R)$. If B is a finitely generated R-module, prove that

$$\bigcap_{n \geq 1} J^n B = \{0\}.$$

Conclude that if $(R, \mathfrak{m})$ is a noetherian local ring, then $\bigcap_{n \geq 1} \mathfrak{m}^n B = \{0\}$.
Hint. Let $D = \bigcap_{n \geq 1} J^n B$, observe that $JD = D$, and use Nakayama's lemma.

11.81 Use the Rees lemma to prove a weaker version of Proposition 11.175: If $(R, \mathfrak{m})$ is a regular local ring, then $D(R) \geq \mu(\mathfrak{m}) = \dim(R)$.
Hint. If $\mathrm{Ext}^d_R(k, R) \neq \{0\}$, then $D(R) > d - 1$; that is, $D(R) \geq d$.
Let $\mathfrak{m} = (x_1, \ldots, x_d)$, where $x_1, \ldots, x_d$ is an R-sequence. Then

$$\begin{aligned}
\mathrm{Ext}^d_R(k, R) &\cong \mathrm{Ext}^{d-1}_{R/(x_1)}(k, R/(x_1)) \\
&\cong \mathrm{Ext}^{d-2}_{R/(x_1, x_2)}(k, R/(x_1, x_2)) \cong \cdots \\
&\cong \mathrm{Ext}^0_k(k, k) \cong \mathrm{Hom}_k(k, k) \cong k \neq \{0\}.
\end{aligned}$$

11.82 Let R be a commutative ring, let M be a finitely generated R-module, and let I be an ideal such that $IM \neq M$.

(i) If $x_1, x_2, \ldots, x_n$ is an M-sequence contained in I, prove that

$$(x_1) \subsetneq (x_1, x_2) \subsetneq \cdots \subsetneq (x_1, \ldots, x_n).$$

(ii) If R is noetherian, prove that there is a longest M-sequence contained in I.

11.83 If k is a field, prove that $k[[x_1, \ldots, x_n]]$ is noetherian.
Hint. Define the *order* $o(f)$ of a nonzero formal power series $f = (f_0, f_1, f_2, \ldots)$ to be the smallest n with $f_n \neq 0$. Find a proof similar to that of the Hilbert basis theorem. (See Zariski-Samuel, *Commutative Algebra* II, page 138.)

11.84 If k is a field, prove that the ring of formal power series $k[[x_1, \ldots, x_n]]$ is a UFD.

Appendix

The Axiom of Choice and Zorn's Lemma

Nowadays, most mathematicians accept the axiomatization **ZFC** of set theory, due to E. Zermelo and A. Fraenkel; the letter C abbreviates *choice*. Using consequences of this axiomatization, we will prove the equivalence of the axiom of choice, the well-ordering principle, and Zorn's lemma. Let us begin by recalling some definitions from Chapter 6.

Definition. If A is a set, let $\mathcal{P}(A)^{\#}$ denote the family of all its nonempty subsets. The ***axiom of choice*** states that if A is a nonempty set, then there exists a function $\beta : \mathcal{P}(A)^{\#} \to A$ with $\beta(S) \in S$ for every nonempty subset S of A. Such a function β is called a ***choice function***.

Informally, the axiom of choice is a harmless looking statement; it says that we can simultaneously choose one element from each nonempty subset of a set. We now show that the axiom of choice is equivalent to a statement we would hate to be false.

Proposition A.1. *The axiom of choice holds if and only if the cartesian product $\prod_{i \in I} X_i$ of nonempty sets is itself nonempty.*[1]

Proof. Let us assume the axiom of choice. Recall that an element of $\prod_{i \in I} X_i$ is an I-tuple $x = (x_i)$ with $x_i \in X_i$ for all $i \in I$. Now an I-tuple is really a function

$$f : I \to A = \bigcup_{i \in I} X_i$$

with $f(i) = x_i \in X_i$ for all $i \in I$. Define $\varphi : I \to \mathcal{P}(A)^{\#}$ by $\varphi(i) = X_i$. If $\beta : \mathcal{P}(A)^{\#} \to A$ is a choice function, then the composite $f = \beta \circ \varphi : I \to A$ satisfies $f(i) = \beta(\varphi(i)) = \beta(X_i) \in X_i$, and so it is an element of $\prod_{i \in I} X_i$. Therefore, the cartesian product is nonempty.

Conversely, let A be a nonempty set. Define $I = \mathcal{P}(A)^{\#}$, and consider $\prod_{S \in I} S$. By hypothesis, this product is nonempty, and so it contains an element β, where $\beta(S) \in S$

[1]By definition, a set X is nonempty if there is an element $x \in X$. If $X_1 \neq \varnothing$ and $X_2 \neq \varnothing$, then there is $x_1 \in X_1$ and $x_2 \in X_2$, and hence $(x_1, x_2) \in X_1 \times X_2$; that is, $X_1 \times X_2 \neq \varnothing$. More generally, we can prove, by induction, that if the index set $I = \{1, 2, \ldots, n\}$ is finite, then $\prod_{i \in I} X_i = X_1 \times \cdots \times X_n \neq \varnothing$. Thus, the axiom of choice is significant only when the index set I is infinite.

for all $S \in \mathcal{P}(A)^{\#}$; that is, β is a choice function for A. Therefore, the axiom of choice holds. •

There are various equivalent forms of the axiom of choice that are more convenient to use, the most popular of which are the *well-ordering principle* and *Zorn's lemma*, which we state after some preliminary definitions. Most mathematicians accept the axiom of choice (as do we), and so they also accept these equivalent forms as well.

Recall that a set X is a *partially ordered set* if there is a relation $x \preceq y$ defined on X that is reflexive, antisymmetric, and transitive.

If it is necessary to display the ordering relation, we may also say that $(X, \preceq)$ is a partially ordered set.

Definition. A partially ordered set X is ***well-ordered*** if every nonempty subset S of X contains a ***smallest element***; that is, there is $s_0 \in S$ with

$$s_0 \preceq s \text{ for all } s \in S.$$

A partially ordered set X is a ***chain*** if any two elements are comparable; that is, for all $x, y \in X$, either $x \preceq y$ or $y \preceq x$.

Example A.2.

(i) The least integer axiom in Chapter 1 says that $\mathbb{N}$, the natural numbers, is well-ordered. More generally, $\mathbb{N}^n$ equipped with a monomial order, defined in Chapter 6, is a well-ordered set.

(ii) The empty set $\varnothing$ is well-ordered; otherwise, $\varnothing$ would contain a nonempty subset (without a smallest element), and this is a contradiction.

(iii) The integers $\mathbb{Z}$ is not well-ordered, for there is no smallest integer.

(iv) The subset X of $\mathbb{Q}$, defined by

$$X = \{1 - \tfrac{1}{n} : n \geq 1\} \,\cup\, \{2 - \tfrac{1}{n} : n \geq 1\}$$

is well-ordered. Note that $1 = 2 - \frac{1}{1}$ has infinitely many predecessors.

(v) Let X be a well-ordered set. An element $\tau \in X$ is a ***top element*** if there is no $\alpha \in X$ with $\tau \prec \alpha$. If $\alpha \in X$ is not the top element of X (should one exist), then $X^\alpha = \{\beta \in X : \alpha \prec \beta\} \neq \varnothing$, and so it has a smallest element α', called the ***successor*** of α. The successor α' is the "next" element after α: formally, $\alpha \prec \alpha'$ and there is no $\beta \in X$ with $\alpha \prec \beta \prec \alpha'$ (if there were such a β, then $\beta \in X^\alpha$ and so $\alpha' \preceq \beta$). An element $\beta \in X$ is a ***limit*** if it is not a successor; that is, there is no $\alpha \in X$ with $\beta = \alpha'$. The smallest element is X is a limit; in part (iv), we saw that $X = \{1 - \frac{1}{n} : n \geq 1\} \,\cup\, \{2 - \frac{1}{n} : n \geq 1\}$ is well-ordered, and it is clear that $1 = 2 - \frac{1}{1}$ is a limit in X. Thus, every element in X is either a successor or a limit. ◂

Here are some basic properties of well-ordered sets.

Proposition A.3.

(i) *Every subset Y of a well-ordered set X is itself well-ordered.*

(ii) *Let X be a well-ordered set. If $x, y \in X$, then either $x \preceq y$ or $y \preceq x$.*

(iii) *If X is a well-ordered set, then every strictly decreasing sequence $x_1 \succ x_2 \succ \cdots$ in X is finite.*

(iv) *Assuming the axiom of choice, the converse of part* (iii) *is true. If X is a chain in which every strictly decreasing sequence $x_1 \succ x_2 \succ \cdots$ in X is finite, then X is well-ordered.*

Proof. (i) If S is a nonempty subset of Y, then it is also a subset of X and, as any nonempty subset of X, it contains a smallest element. Therefore, Y is well-ordered.

(ii) The subset $S = \{x, y\}$ has a smallest element, which is either x or y. In the first case, $x \preceq y$, and in the second case, $y \preceq x$.

(iii) If X is well-ordered, then $S = \{x_1, x_2, \ldots\}$ has a smallest element, say, x_i; that is, $x_n \succeq x_i$ for all $n \geq 1$. In particular, if $n = i + 1$, then $x_{i+1} \succeq x_i$, which contradicts $x_i \succ x_{i+1}$.

(iv) Assume that there exists a nonempty subset S of X that has no smallest element. Choose $s_0 \in S$; since s_0 is not smallest, it is not true that $s_0 \preceq s$ for all $s \in S$. Thus, either there exists $s_1 \in S$ with $s_0 \succ s_1$ or there is $s \in S$ with s_0 and s not comparable; the latter cannot occur because X is a chain. Similarly, there is $s_2 \in S$ with $s_1 \succ s_2$. By induction, for all $n \geq 0$, there are elements $s_i \in S$ with $s_0 \succ s_1 \cdots \succ s_n \succ s_{n+1}$. We want to assemble these infinitely many choices, one for each n, into one descending sequence[2]; that is, we want a function $f : \mathbb{N} \to S$ with $f(n) = s_n$. Here is the formal way to do this. Let $\mathcal{F}$ be the family of all functions g from all initial segments $\{0, 1, \ldots, n\} \to S$, and let β be a choice function on $\mathcal{F}$: that is, $\beta(T) \in T$ for every nonempty subset $T \subseteq \mathcal{F}$. We use β to construct the desired sequence. Choose an element $s_0 \in S$, which is possible because $S \neq \varnothing$. Define $F_0 = \{g \in \mathcal{F} : \text{domain}(g) = \{0\} \text{ and } g(0) = s_0\})$ (there is only one g in F_0), and define $g_0 = \beta(F_0)$. For $n > 0$, we know, by induction, that $F_{n+1} \neq \varnothing$, where

$$F_{n+1} = \big\{g \colon \{0, 1, \ldots, n+1\} \to X : g\big|\{0, \ldots, n\} = g_n \text{ and } g(n) \succ g(n+1)\big\}.$$

Therefore, we may define $g_{n+1} = \beta(F_{n+1})$. Finally, define g^* to be the union of the g_n; that is, $g^*(n) = g_n(n)$ for all n. The function g^* is a strictly descending sequence in S, and this contradicts the hypothesis that every strictly decreasing sequence in S is finite. •

Well-ordering Principle. *Every set X has some well-ordering of its elements.*

---

[2]We have already done this, without comment, in Proposition 6.38, when we showed that ACC implies the maximum condition. Actually, the proof only uses a weaker form of the axiom of choice in which the index set is countable.

If X happens to be a partially ordered set, then a well-ordering, whose existence is asserted by the well-ordering principle, may have nothing to do with the original partial ordering. For example, $\mathbb{Z}$ can be well-ordered:

$$0 \preceq 1 \preceq -1 \preceq 2 \preceq -2 \preceq \cdots .$$

That $\mathbb{N}$ is well-ordered is just another way of stating mathematical induction. Thus, the well-ordering principle suggests the possibility of a generalized induction that applies to a collection of statements indexed by a well-ordered set of any, possibly uncountable, cardinality. Such a generalization does, in fact, exist, and it is called ***transfinite induction***. Let $\{S(\alpha) : \alpha \in I\}$ be a family of statements indexed by a well-ordered set I. If α_0 is the smallest index in I, then the *base step* is the statement that $S(\alpha_0)$ is true. The *inductive step* is the statement that if β is an index and $S(\alpha)$ is true for all $\alpha \prec \beta$, then $S(\beta)$ is true. Transfinite induction says that if the base step and the inductive step hold, then all the statements $S(\alpha)$ are true. (Often, the proof of the inductive step splits into two cases, depending on whether β is a successor or a limit). Here is a surprising use of transfinite induction: There exists a subset Q of the plane that intersects every straight line in exactly two points. The idea of the proof is to construct Q by well-ordering the set of all the lines in the plane in such a way that every line has only countably many predecessors. Now, from each line in turn, judiciously select at most two of its points to put into Q.

We will be able to state Zorn's lemma after the following definitions.

Definition. Let X be a partially ordered set. An ***upper bound*** of a subset S of X is an element $x \in X$, not necessarily in S, such that

$$s \preceq x \text{ for all } s \in S.$$

An element $m \in X$ is a ***maximal element*** if there is no $x \in X$ for which $m \prec x$; that is, if $x \in X$ and if $m \preceq x$, then $m = x$.

A partially ordered set may have no maximal elements: for example, $\mathbb{R}$, with its usual ordering, is a chain having no maximal elements. A partially ordered set may have many maximal elements: for example, if X is the partially ordered set of all the proper subsets of a set U, then a subset S is a maximal element if and only if $S = U - \{u\}$ for some $u \in U$; that is, S is the complement of a point.

Zorn's lemma is a criterion that guarantees the existence of maximal elements.

Zorn's lemma. *If X is a nonempty partially ordered set in which every chain has an upper bound, then X has a maximal element.*

Theorem A.4. *The following statements are equivalent.*

(i) *Zorn's lemma.*

(ii) *The well-ordering principle.*

(iii) *The axiom of choice.*

We split Theorem A.4 into three separate theorems. Let us begin with a definition and a lemma.

Definition. If X is a well-ordered set and $c \in X$, then the ***open segment*** $\text{Seg}(c)$ is the subset

$$\text{Seg}(c) = \{x \in X : x \prec c\}.$$

The next result supplements Proposition A.3.

Lemma A.5. *A chain X is well-ordered if and only if every open segment of X is well-ordered.*

Proof. Necessity is obvious, for every subset of a well-ordered set is well-ordered. Conversely, let S be a nonempty subset of X. Of course, if S is a singleton, then it contains a smallest element, and so we may assume that S contains at least two elements, say, c' and c. Since X is a chain, we may assume that $c' \prec c$. Hence, $\text{Seg}(c) \cap S \neq \varnothing$; as every nonempty subset of a well-ordered set is well-ordered, there is a smallest element, say, z, in $\text{Seg}(c) \cap S$. Now z is the smallest element in S, for if there is $s' \in S$ with $s' \prec z$, then $s' \in \text{Seg}(c) \cap S$, contradicting z being the smallest element in $\text{Seg}(c) \cap S$. Therefore, X is well-ordered. •

In general, an ascending union of well-ordered subsets of a partially ordered set need not be well-ordered. For example, it is easy to see that for every positive integer n, the subset

$$S_n = \{m \in \mathbb{Z} : m \geq -n\}$$

is a well-ordered subset of $\mathbb{Z}$, but $\bigcup_n S_n = \mathbb{Z}$ is not well-ordered.

With an extra assumption, we can force a union of well-ordered subsets to be well-ordered.

Notation. If B and C are subsets of a partially ordered set X, then we write

$$B \trianglelefteq C$$

if either $B = C$ or B is an open segment of C; that is, there exists $c \in C$ with $B = \text{Seg}(c)$.

Lemma A.6. *Let $(X, \preceq)$ be a partially ordered set, and let $\{S_i : i \in I\}$ be a family of well-ordered subsets of X indexed by some set I. If, for each i, j, either $S_i \trianglelefteq S_j$ or $S_j \trianglelefteq S_i$, then $\bigcup_{i \in I} S_i$ is a well-ordered subset of X.*

Proof. Let $U = \bigcup_i S_i$. By Lemma A.5, it suffices to show that any open segment $\text{Seg}(c)$, where $c \in U$, is well-ordered. Now $c \in S_i$ for some i; since S_i is well-ordered, so is any of its subsets; thus, it suffices to show that $\text{Seg}(c) = \{x \in U : u \prec c\} \subseteq S_i$; that is, if $u \prec c$, we must show that $u \in S_i$. Now $u \in S_j$ for some j. If $S_j \trianglelefteq S_i$, then $u \in S_i$ and we are done. If $S_i \trianglelefteq S_j$, then $S_i \subseteq S_j$, so that $c \in S_j$; moreover, since S_i is an open segment of S_j, $u \prec c$ implies $u \in S_i$, as desired. •

Definition. A subset A of a well-ordered set $(X, \preceq)$ is ***closed*** in X if $A \neq \varnothing$ and if $x \preceq a$, where $x \in X$ and $a \in A$, implies $x \in A$. (Thus, if A is closed and $a \in A$, then A contains everything smaller than a as well.)

Given a well-ordered set X and $c \in X$, it is obvious that the "closed segment"

$$A = \{x \in X : x \preceq c\}$$

is a closed subset. If c is the top element of X (should such exist), then $A = X$; if c is not a top element, then it has a successor c', and $A = \mathrm{Seg}(c')$. Thus, closed segments are closed subsets, but they are nothing new.

Lemma A.7. *If $(X, \preceq)$ is a well-ordered set, then A is closed in X if and only if $A \trianglelefteq X$; that is, either $A = X$ or $A = \mathrm{Seg}(c)$ for some $c \in X$.*

Proof. It is clear that open segments are closed, and so only necessity needs proof.

Assume that A is closed; we must show that if $A \neq X$, then A is an open segment. Since X is well-ordered, there is a smallest element $c \in X - A$, and we claim that $A = \mathrm{Seg}(c)$. If there is some $a \in A$ with $c \preceq a$, then $c \in A$, because A is closed, and this contradicts $c \notin A$. Therefore, $a \prec c$ for all $a \in A$ (we are using the fact that well-ordered sets are chains); that is, c is an upper bound of A, and so $A \subseteq \mathrm{Seg}(c)$. For the reverse inclusion, suppose that $x \in \mathrm{Seg}(c)$; that is, $x \prec c$. If $x \notin A$, then $x \in X - A$, and so $c \preceq x$, a contradiction. •

Here is the first step of Theorem A.4. .

Theorem A.8. *If Zorn's lemma holds, then the well-ordering principle holds*: *every set X can be well-ordered.*

Proof. Since $\varnothing$ is well-ordered, we may assume that $X \neq \varnothing$. Let $\mathcal{L}$ be the family of all the well-ordered subsets of X; more precisely, an element of $\mathcal{L}$ is an ordered pair $(S, \sqsubseteq)$ consisting of a subset S of X together with some well-ordering $\sqsubseteq$ of it. Thus, a subset S of X may appear several times in $\mathcal{L}$, equipped with different well-orderings. We now make $\mathcal{L}$ into a partially ordered set. Define

$$(S, \sqsubseteq) \preceq (S', \sqsubseteq')$$

to mean either (i) $S = S'$ and $\sqsubseteq = \sqsubseteq'$ or (ii) $S \subsetneq S'$, the orderings coincide on S (that is, if $a, b \in S$, then $a \sqsubseteq b$ holds if and only if $a \sqsubseteq' b$ holds), and S is an open segment of S'.

We now show that $\mathcal{L}$ satisfies the hypothesis of Zorn's lemma. Note that $\mathcal{L} \neq \varnothing$, for any 1-point subset is a well-ordered subset, and so it gives an element of $\mathcal{L}$. Let $C = \{(S_i, \preceq_i) : i \in I\}$ be a chain in $\mathcal{L}$; that is, for each i, j, either $(S_i, \sqsubseteq_i) \preceq (S_j, \sqsubseteq_j)$ or $(S_j, \sqsubseteq_j) \preceq (S_i, \sqsubseteq_i)$. Define $U = \bigcup_i S_i$, and define a partial order $\sqsubseteq$ on U as follows. If $u, v \in U$, then there are indices i and j with $u \in S_i$ and $v \in S_j$; we may assume that $(S_i, \sqsubseteq_i) \preceq (S_j, \sqsubseteq_j)$, so that both $u, v \in S_j$, and we define $u \sqsubseteq v$ if $u \sqsubseteq_j v$. This definition does not depend on the choice of indices, for if there are indices k and ℓ with $u \in S_k$ and $v \in S_\ell$, then $(S_k, \sqsubseteq_k) \preceq (S_\ell, \sqsubseteq_\ell)$, say, and $u \sqsubseteq_\ell v$ is a competing definition.

But $(S_j, \sqsubseteq_j) \preceq (S_\ell, \sqsubseteq_\ell)$, and so $u \sqsubseteq_j v$ if and only if $u \sqsubseteq_\ell v$. It is now routine to prove that $(U, \sqsubseteq)$ is a partially ordered set. In fact, $(U, \sqsubseteq)$ is well-ordered, by Lemma A.6, and so $(U, \sqsubseteq) \in \mathcal{L}$. Furthermore, we claim that each $(S_i, \sqsubseteq_i)$ is closed in $(U, \sqsubseteq)$. Suppose that $u \sqsubseteq s_i$, where $s_i \in S_i$ and $u \in U$. Now $u \in S_j$ for some j; if $(S_j, \sqsubseteq_j) \preceq (S_i, \sqsubseteq_i)$, then $u \in S_i$ as desired; if $(S_i, \sqsubseteq_i) \preceq (S_j, \sqsubseteq_j)$, then S_i is closed in S_j, and so $u \in S_i$. Lemma A.7 now gives $(S_i, \sqsubseteq_i) \preceq (U, \sqsubseteq)$ for all i; that is, $(U, \sqsubseteq)$ is an upper bound of C.

By Zorn's lemma, $\mathcal{L}$ has a maximal element, say $(M, \leq)$. If M contains every element of X, then X can be well-ordered. If there is some $x \in X$ with $x \notin M$, then define a well-ordering $\leq'$ of $M \cup \{x\}$ extending the given well-ordering of M by defining $m <' x$ for every $m \in M$. Since $M = \mathrm{Seg}(x)$ in $M \cup \{x\}$, we have $(M, \leq) \prec (M \cup \{x\}, \leq')$, contradicting the maximality of $(M, \leq)$. Therefore, $M = X$ and X can be well-ordered. •

Almost all proofs involving Zorn's lemma have the same format: define an appropriate nonempty partially ordered set; show that its chains have upper bounds; show that a maximal element, whose existence is guaranteed by Zorn's lemma, can be used to prove the theorem (this last step is usually an indirect proof).

Here is a small comment about the axiom of choice before we present the second step in the proof of Theorem A.4. Given sets A and X, one way to define a function $f : A \to X$ is to specify its values. For example, there exists a function $f : \mathbb{N} \to \mathbb{N}$ with $f(n) = n+1$ for each $n \in \mathbb{N}$. Not every function is given by a formula, however, and the axiom of choice deals with the problem of when a function is actually defined. If $\{G_a : a \in A\}$ is a family of groups, then we may define a choice function $f : A \to \bigcup_a G_a$ by $f(a) = 1_a$, where 1_a is the identity element of G_a; we do not need the axiom of choice to define f. In contrast, if we merely "choose" some element $x_a \in G_a$, then the "function" $h : A \to \bigcup_a G_a$ with $h(a) = x_a$ is not well-defined. Such a "function" h ought not be a bona fide function. How could one possibly detect any properties of h; for example, is h an injection?

Theorem A.9. *The well-ordering principle implies the axiom of choice.*

Proof. Let A be a nonempty set. We may assume that A has some well-ordering of its elements, and it follows that every nonempty subset of A is also well-ordered. Define a choice function $\beta : \mathcal{P}(A)^\# \to A$ by defining, for each nonempty subset S of A, $\beta(S)$ to be the smallest element of S. •

Daniel Grayson has shown me an elegant proof that the axiom of choice implies Zorn's lemma; it is a variant of a proof of E. Zermelo in 1904, as adapted by H. Kneser in 1950.

Theorem A.10. *The axiom of choice implies Zorn's lemma.*

Proof. Assume that X has no maximal elements. If A is a well-ordered subset of X, then A is a chain, and hence A has an upper bound, say x. Since x is not a maximal element, there exists $y \in X$ with $x \prec y$; it follows that every well-ordered subset A has an upper bound that is not in A. Let $\mathcal{W}$ denote the family of all the well-ordered subsets of X. For each $A \in \mathcal{W}$, define

$$U_A = \{\text{all upper bounds } u \text{ of } A \text{ with } u \notin A\};$$

each $U_A \neq \varnothing$, by hypothesis, and so Proposition A.1 says there is some g in $\prod_{A \in \mathcal{W}} U_A$. Thus, for all $A \in \mathcal{W}$, we have $g(A)$ an upper bound of A and $g(A) \notin A$.

We use g to construct some special well-ordered subsets. Define an element $c_0 \in X$ by $c_0 = g(\varnothing)$. Call a well-ordered subset C of X a ***g-set*** if c is the upper bound of $C \cap \text{Seg}(c)$ chosen by g;[3] that is, $c_0 \in C$ and $c = g(C \cap \text{Seg}(c))$ for every $c \in C$.

We are going to show that the union of all the g-sets is itself a g-set, and this will then be shown to give a contradiction.

If C and D are g-sets, we claim that either $C \trianglelefteq D$ or $D \trianglelefteq C$. Define W to be the union of all those subsets B with $B \trianglelefteq C$ and $B \trianglelefteq D$. We claim that $W \trianglelefteq C$ and $W \trianglelefteq D$; that is, W is closed in C and in D. Take $w \in W$; this element got into W because it lies in some B, where $B \trianglelefteq C$ and $B \trianglelefteq D$. If $c \in C$ and $c \preceq w$, then $c \in B$ (because B is closed in C). Hence, $c \in B \subseteq W$ (for W is, by definition, the union of all such subsets B). Therefore, W is closed in C. Similarly, W is closed in D. If either $W = C$ or $W = D$, then the claim is true. Hence, we may assume that $W \triangleleft C$ [so that $W = C \cap \text{Seg}(c')$ for some $c' \in C - W$], and $W \triangleleft D$ [so that $W = D \cap \text{Seg}(d')$ for some $d' \in D - W$]. Since C and D are g-sets, $c' = g(C \cap \text{Seg}(c')) = g(W)$ and $d' = g(D \cap \text{Seg}(d')) = g(W)$. Therefore, $c' = d'$. But now $W \cup \{c'\} = W \cup \{c'\}$ is closed in C and in D, for it is a closed interval. Thus, $W \cup \{c'\} \subseteq W$, contradicting $c' \notin W$. Therefore, either $W = C$ or $W = D$; that is, either $C \trianglelefteq D$ or $D \trianglelefteq C$, as claimed.

Finally, let Ω be the union of all the g-sets. The just-established claim shows that the hypothesis of Lemma A.6 is satisfied, and so Ω is a well-ordered subset. Let us show that Ω is itself a g-set. If $c \in \Omega$, then there is some g-set C containing c, and $c = g(C \cap \text{Seg}(c))$. But $C \trianglelefteq \Omega$, by Lemma A.7 and the fact just proved above that either $C \trianglelefteq \Omega$ or $\Omega \trianglelefteq C$; hence, $C \cap \text{Seg}(c) = \Omega \cap \text{Seg}(c)$. Therefore, $c = g(\Omega \cap \text{Seg}(c))$, and Ω is a g-set. On the other hand, $\Omega' = \Omega \cup \{g(\Omega)\}$ is a g-set not contained in Ω, and this is a contradiction. We conclude that no such function g can exist, and hence that X has a maximal element. •

It appears that we have used a weaker hypothesis than that of Zorn's lemma: only well-ordered subsets need upper bounds. However, it is shown in Exercise 6.45 on page 374 that every chain C in a partially ordered set contains a well-ordered subset W such that C and W have the same upper bounds. Hence, if all well-ordered subsets have upper bounds, then all chains have upper bounds as well.

If $B = C$, then $W = C$; otherwise, B is an open segment of C, and so there is $b \in C$ with $B = \text{Seg}(b)$. If B' is also a proper subset of C, then $B' = \text{Seg}(b')$. We may assume that $b \preceq b'$, and so $B \trianglelefteq B'$. It now follows from Lemma A.7 that $W \trianglelefteq C$.

[3] Each of the following sets are g-sets. If $c_1 = g(\{c_0\})$, define $c_2 = g(\{c_0, c_1\})$, and, by induction, $c_{n+1} = g(\{c_0, \ldots, c_n\})$. Note that $c_0 \prec c_1 \prec c_2 \prec \cdots$. Each subset $\{c_0, c_1, \ldots, c_n\}$ is a g-set. There are infinite g-sets as well. For example, if $C' = \{c_n : n \in \mathbb{N}\}$, let $c' = g(C')$, and define $C'' = C' \cup \{c'\}$.

Bibliography

Adem, A., and Milgram, R. J., *Cohomology of Finite Groups*, Springer–Verlag, Berlin, 1994.

Albert, A. A., editor, *Studies in Modern Algebra*, MAA Studies in Mathematics, vol. 2, Mathematical Association of America, Washington, 1963.

Artin, E., *Geometric Algebra*, Interscience Publishers, New York, 1957.

Artin, E., Nesbitt, C. J., and Thrall, R. M., *Rings with Minimum Condition*, University of Michigan Press, Ann Arbor, 1968.

Aschbacher, M., *Finite Group Theory*, Cambridge University Press, Cambridge, 1986.

Atiyah, M., and Macdonald, I. G., *Introduction to Commutative Algebra*, Addison–Wesley, Reading, 1969.

Biggs, N. L., *Discrete Mathematics*, Oxford University Press, 1989.

Birkhoff, G., and Mac Lane, S., *A Survey of Modern Algebra*, 4th ed., Macmillan, New York, 1977.

Blyth, T. S., *Module Theory; an Approach to Linear Algebra*, Oxford University Press, 1990.

Borevich, Z. I., and Shafarevich, I. R., *Number Theory*, Academic Press, Orlando, 1966.

Bourbaki, N., *Elements of Mathematics*; *Algebra* I; *Chapters* 1-3, Springer–Verlag, New York, 1989.

————, *Elements of Mathematics*; *Commutative Algebra*, Addison–Wesley, Reading, 1972.

Brown, K. S., *Cohomology of Groups*, Springer–Verlag, Berlin, 1982.

Bruns, W., and Herzog, J., *Cohen–Macaulay Rings*, Cambridge University Press, 1993.

Buchberger, B., and Winkler, F., editors, *Gröbner Bases and Applications*, LMS Lecture Note Series 251, Cambridge University Press, 1998.

Burnside, W., *The Theory of Groups of Finite Order*, 2d ed., Cambridge University Press, 1911; Dover reprint, Mineola, 1955.

Caenepeel, S., *Brauer Groups, Hopf Algebras, and Galois Theory*, Kluwer, Dordrecht, 1998.

Cajori, F., *A History of Mathematical Notation*, Open Court, 1928; Dover reprint, Mineola, 1993.

Carmichael, R., *An Introduction to the Theory of Groups*, Ginn, New York, 1937.

Carter, R., *Simple Groups of Lie Type*, Cambridge University Press, Cambridge, 1972.

Cassels, J. W. S., and Fröhlich, A., *Algebraic Number Theory*, Thompson Book Co., Washington, D.C., 1967.

Conway, J. H., Curtis, R. T., Norton, S. P., Parker, R. A., Wilson, R. A., *ATLAS of Finite Groups*, Oxford University Press, 1985.

Cox, D., Little, J., and O'Shea, D., *Ideals, Varieties, and Algorithms*, 2d ed., Springer–Verlag, New York, 1997.

Coxeter, H. S. M., and Moser, W. O. J., *Generators and Relations for Discrete Groups*, Springer–Verlag, New York, 1972.

Curtis, C. W., and Reiner, I., *Representation Theory of Finite Groups and Associative Algebras*, Interscience, New York, 1962.

Dieudonné, J., *La Géometrie des Groupes Classiques*, Springer–Verlag, Berlin, 1971.

Dixon, J. D., du Sautoy, M. P. F., Mann, A., and Segal, D., *Analytic Pro-p Groups*, Cambridge University Press, 1991.

Dornhoff, L., *Group Representation Theory, Part* A, *Ordinary Representation Theory*, Marcel Dekker, New York, 1971.

Drozd, Yu. A., and Kirichenko, V. V., *Finite Dimensional Algebras*, Springer–Verlag, New York, 1994.

Dummit, D. S., and Foote, R. M., *Abstract Algebra*, 2nd ed., Prentice Hall, Upper Saddle River, 1999.

Eisenbud, D., *Commutative Algebra with a View Toward Algebraic Geometry*, Springer–Verlag, New York, 1995.

Evens, L., *The Cohomology of Groups*, Oxford Mathematical Monographs, Oxford University Press, New York, 1991.

Farb, B., and Dennis, R. K., *Noncommutative Algebra*, Springer–Verlag, New York, 1993.

Feit, W., *Characters of Finite Groups*, W. A. Benjamin, New York, 1967.

Fröhlich, A., and Taylor, M. J., *Algebraic Number Theory*, Cambridge Studies in Advanced Mathematics 27, Cambridge University Press, 1991.

Fuchs, L., *Infinite Abelian Groups* I, Academic Press, Orlando, 1970.

———, *Infinite Abelian Groups* II, Academic Press, Orlando, 1973.

Fulton, W., *Algebraic Curves*, Benjamin, New York, 1969.

———-, *Algebraic Topology*; *A First Course*, Springer–Verlag, New York, 1995.

Gaal, L., *Classical Galois Theory with Examples*, 4th ed., Chelsea, American Mathematical Society, Providence, 1998.

Gorenstein, D., Lyons, R. and Solomon, R., *The Classification of the Finite Simple Groups*, Math. Surveys and Monographs Volume 40, American Mathematical Society, Providence, 1994.

Greub, W. H., *Multilinear Algebra*, Springer–Verlag, New York, 1967.

Hadlock, C., *Field Theory and Its Classical Problems*, Carus Mathematical Monographs, Mathematical Association of America, Washington, 1978.

Hahn, A. J., *Quadratic Algebras, Clifford Algebras, and Arithmetic Witt Groups*, Universitext, Springer–Verlag, New York, 1994.

Hardy, G. H., and Wright, E. M., *An Introduction to the Theory of Numbers*, 4th ed., Oxford University Press, 1960.

Harris, J., *Algebraic Geometry*, Springer–Verlag, New York, 1992.

Hartshorne, R., *Algebraic Geometry*, Springer–Verlag, New York, 1977.

Herstein, I. N., *Topics in Algebra*, 2d ed., Wiley, New York, 1975.

———, *Noncommutative Rings*, Carus Mathematical Monographs No. 15, Mathematical Association of America, Washington, 1968.

Humphreys, J. E., *Introduction to Lie Algebras and Representation Theory*, Springer–Verlag, New York, 1972.

Huppert, B., *Character Theory of Finite Groups*, de Gruyter, Berlin, 1998.

———, *Endliche Gruppen* I, Springer–Verlag, New York, 1967.

Isaacs, I. M., *Character Theory of Finite Groups*, Academic Press, San Diego, 1976.

———-, *Algebra, A Graduate Course*, Brooks/Cole Publishing, Pacific Grove, 1994.

Jacobson, N., *Basic Algebra* I, Freeman, San Francisco, 1974.

———, *Basic Algebra* II, Freeman, San Francisco, 1980.

———, *Finite-Dimensional Division Algebras over Fields*, Springer–Verlag, New York, 1996.

———, *Lie Algebras*, Interscience Tracts Number 10, Wiley, New York, 1962.

———, *Structure of Rings*, Colloquium Publications 37, American Mathematical Society, Providence, 1956.

Kaplansky, I., *Commutative Rings*, University of Chicago Press, 1974.

————–, *Fields and Rings*, 2d ed., University of Chicago Press, 1972.

————–, *Infinite Abelian Groups*, University of Michigan Press, Ann Arbor, 1969.

————–, *Linear Algebra and Geometry*; *a Second Course*, Allyn & Bacon, Boston, 1969.

————–, *Set Theory and Metric Spaces*, Chelsea, American Mathematical Society, Providence, 1977.

Kostrikin, A. I., and Shafarevich, I. R. (editors), *Encyclopaedia of Mathematical Sciences*, *Algebra* IX: *Finite Groups of Lie Type*; *Finite-Dimensional Division Algebras*, Springer–Verlag, New York, 1996.

Lam, T. Y., *The Algebraic Theory of Quadratic Forms*, Benjamin, Reading, 1973, 2d. revised printing, 1980.

———–, *A First Course in Noncommutative Rings*, Springer–Verlag, New York, 1991.

———–, *Lectures on Modules and Rings*, Springer–Verlag, New York, 1999.

Lang, S., *Algebra*, 3d ed., Addison–Wesley, Reading, 1993.

Lidl, R., and Niederreiter, H., *Introduction to Finite Fields and Their Applications*, University Press, Cambridge, 1986.

Lyndon, R. C., and Schupp, P. E., *Combinatorial Group Theory*, Springer–Verlag, New York, 1977.

Macdonald, I. G., *Algebraic Geometry*; *Introduction to Schemes*, Benjamin, New York, 1968.

Mac Lane, S., *Categories for the Working Mathematician*, Springer–Verlag, New York, 1971.

————, *Homology*, Springer–Verlag, New York, 3d corrected printing, 1975.

Mac Lane, S., and Birkhoff, G., *Algebra*, MacMillan, New York, 1967.

Malle, G., and Matzat, B., *Inverse Galois Theory*, Springer–Verlag, New York, 1999.

Matsumura, H., *Commutative Ring Theory*, Cambridge University Press, 1986.

McCleary, J., *User's Guide to Spectral Sequences*, Publish or Perish, Wilmington, 1985.

McConnell, J. C., and Robson, J. C., *Noncommutative Noetherian Rings*, Wiley, New York, 1987.

McCoy, N. H., and Janusz, G. J., *Introduction to Modern Algebra*, 5th ed., Wm. C. Brown Publishers, Dubuque, Iowa, 1992.

Milnor, J., *Introduction to Algebraic K-Theory*, Annals of Mathematical Studies, No. 72, Princeton University Press, 1971.

Montgomery, S. and Ralston, E. W., *Selected Papers on Algebra*, Raymond W. Brink Selected Mathematical Papers, volume 3, Mathematical Association of America, Washington, 1977.

Mumford, D., *The Red Book of Varieties and Schemes*, Lecture Notes in Mathematics 1358, Springer–Verlag, New York, 1988.

Neukirch, J., Schmidt, A., and Wingberg, K., *Cohomology of Number Fields*, Grundlehren der mathematicshcen Wissenschaften, vol. 323, Springer–Verlag, New York, 2000.

Niven, I., and Zuckerman, H. S., *An Introduction to the Theory of Numbers*, Wiley, New York, 1972.

Northcott, D. G., *Ideal Theory*, Cambridge University Press, 1953.

O'Meara, O. T., *Introduction to Quadratic Forms*, Springer–Verlag, New York, 1971.

Orzech, M., and Small, C., *The Brauer Group of Commutative Rings*, Lecture Notes in Pure and Applied Mathematics, Marcel Dekker, New York, 1975.

Pollard, H., *The Theory of Algebraic Numbers*, Carus Mathematical Monographs Number 9, Mathematical Association of America, 1950.

Procesi, C., *Rings with Polynomial Identities*, Marcel Dekker, New York, 1973.

Reiner, I., *Maximal Orders*, Academic Press, London, 1975; Oxford University Press, 2003.

Robinson, D. J. S., *A Course in the Theory of Groups*, 2d ed., Springer–Verlag, New York, 1996.

Rosenberg, J., *Algebraic K-Theory and Its Applications*, Springer–Verlag, New York, 1994.

Rotman, J. J., *A First Course in Abstract Algebra*, 2d ed., Prentice Hall, Upper Saddle River, 2000.

————–, *Galois Theory*, 2d ed., Springer–Verlag, New York, 1998.

————–, *An Introduction to Homological Algebra*, Academic Press, Orlando, 1979.

————–, *An Introduction to the Theory of Groups*, 4th ed., Springer–Verlag, New York, 1995.

—————, *Journey into Mathematics*, Prentice Hall, Upper Saddle River, 1998.

Rowen, L. H., *Polynomial Identities in Ring Theory*, Academic Press, New York, 1980.

Samuel, P. *Algebraic Theory of Numbers*, Houghton Mifflin, Boston, 1970.

Serre, J.-P., *Algèbre Locale: Multiplicités*, Lecture Notes in Mathematics 11, 3d ed., Springer–Verlag, New York, 1975.

————, *Corps Locaux*, Hermann, Paris, 1968.

————, *Trees*, Springer–Verlag, New York, 1980.

Sims, C. C., *Computation with Finitely Presented Groups*, Cambridge University Press, 1994.

Stillwell, J., *Mathematics and Its History*, Springer–Verlag, New York, 1989.

Suzuki, M., *Group Theory* I, Springer–Verlag, New York, 1982.

Tignol, J.-P., *Galois' Theory of Algebraic Equations*, Wiley, New York, 1988; World Scientific, Singapore, 2001.

van der Waerden, B. L., *Geometry and Algebra in Ancient Civilizations*, Springer–Verlag, New York, 1983.

———————-, *A History of Algebra*, Springer–Verlag, New York, 1985.

———————-, *Modern Algebra*, 4th ed., Ungar, New York, 1966.

———————-, *Science Awakening*, Wiley, New York, 1963.

Weibel, C., *An Introduction to Homological Algebra*, Cambridge University Press, 1994.

Weyl, H., *The Classical Groups; Their Invariants and Representations*, Princeton, 1946.

Weiss, E., *Cohomology of Groups*, Academic Press, Orlando, 1969.

Zariski, O., and Samuel, P., *Commutative Algebra* I, van Nostrand, Princeton, 1958.

————————, *Commutative Algebra* II, van Nostrand, Princeton, 1960.

Index